EXPERIENCING LANGUAGE IN THE ORDER THAT CHILDREN DO: TRAINING ON AGE-ORDERED CHILD-DIRECTED SPEECH FACILITATES SEMANTIC CATEGORY LEARNING IN A RECURRENT NEURAL NETWORK

BY

PHILIP HUEBNER

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Adviser:

Assistant Professor Jon Willits

ABSTRACT

Previous work has shown that semantic category knowledge can be captured by a distributional learning algorithm operating over naturalistic, noisy child-directed speech (Huebner & Willits, 2018). In chapter 1 of this work, I discuss the algorithm behind this study, and its ability to represent hierarchically organized and abstract knowledge. In chapter 2, I replicate the findings of Huebner & Willits (2018) using a variant of their corpus in which fewer post-processing modifications were applied to the raw transcripts. In chapter 3, I investigate whether training on input in order that children actually experience language provides any learning advantage relative to training in the reverse order Indeed, I found that semantic categorization benefits from training on input which was ordered by the age of the target child compared to input which was ordered in reverse. I refer to this effect as the age-order effect. To investigate what corpus-statistical factors may underlie the age-order effect, I explore structural differences between speech to younger vs. older children in chapter 4. In alignment with previous studies, I found that speech to younger children is syntactically less complex compared to speech to older children. Evidence for differences in semantic category structure was inconsistent. In chapter 5, I propose a number of competing explanations of the age-order effect, and identify one hypothesis, termed the good-start hypothesis, as the most promising. In chapter 6, I expand and refine the good-start hypothesis, and provide further empirical support for it. In chapter 7, I test two core assumptions of the theory developed in chapter 6 using carefully controlled artificial language corpora and find strong support for both. I close with a brief overview of findings in infant behavioral studies consistent with the theory and discuss the implications of the theory for infant acquisition of semantic category knowledge.

ACKNOWLEDGEMENTS

I am grateful to my advisor Dr. Jon Willits who opened my eyes to the intricate world of language acquisition research. The moment I arrived at his laboratory, Jon started me on a training regime consisting of classic literature in cognitive science. Our early discussions concerning theoretical debates that are dividing the field, provided me with the kind of inspiration and drive that still motivates me today. Being the first graduate student in his laboratory, there were many logistic, and technological hurdles to figure out; Jon trusted me to make my own decisions with little to no looking-over-my-shoulder, allowing me to develop my own scientific toolbox, and perspective about controversial scientific questions. While an environment of creative freedom has often led me down a rabbit-hole, Jon has a special knack for getting me back on my feet.

TABLE OF CONTENTS

CHAPTER 1: INSTANTIATING THE DISTRIBUTIONAL HYPOTHESIS

The development of semantic memory is an extremely complex phenomenon, requiring input from all perceptual modalities and making use of many psychological processes. Recent research efforts into this topic have focused on the *distributional hypothesis* (Harris, 1954; Firth, 1957), the claim that the similarity, class membership, or relations between linguistic units or concepts can be inferred from the statistical or structural contexts in which those units occur. In the computational realm, this idea was formalized in a range of different models of adult semantics, such as Latent Semantic Analysis (LSA, Landauer and Dumais, 1997), the Hyperspace Analogue to Language (HAL, Lund and Burgess, 1996), Bound Encoding of the Aggregate Language Environment (BEAGLE, Jones and Mewhort, 2007), and Probabilistic Topic Models (Topics, Steyvers and Tenenbaum, 2005). These models use distributional information to construct semantic feature vectors for words. Feature vectors can be composed of concrete associations between words/concepts, as in the HAL model, or they can consist of abstract or latent features that are formed over the course of learning (as in the other three models). The semantic similarity of two words can then be calculated by measuring the similarity of the two words' feature vectors (Kahneman and Tversky, 1972; Smith et al., 1974). Considerable research has since shown that these and related procedures for representing semantic similarity predict a wide range of adult psycholinguistic variables, such as semantic priming and explicit similarity judgments (Burgess and Lund, 1998; Jones et al., 2006; Bullinaria and Levy, 2007; Olney et al., 2012; Pereira et al., 2016).

Concurrent to work in computational modeling of semantic memory, researchers in child language acquisition were studying whether children are sensitive to distributional information,

and whether they can use it to infer word meanings. Gleitman (1990) suggested that syntactic bootstrapping (i.e., inferring aspects of a word's meaning from its syntactic structure) may be an important mechanism by which children begin to learn the meanings of words. Using syntactic bootstrapping, children may, for example, infer whether a verb is transitive or intransitive by tracking whether the verb occurs with one or two nouns or noun phrases. Recent studies have shown that infants and children are sensitive to the distributional structure of words, and do seem to infer aspects of word meaning from lexical and syntactic distributional structure (Fisher et al., 2010; Lany and Saffran, 2010; Syrett and Lidz, 2010; Wojcik and Saffran, 2013).

Instantiating the distributional hypothesis

Thus, both computational and experimental work has shown that substantial semantic information exists in words' distributions, and that human learners are sensitive to this information. But precisely how does a learning system that instantiates the distributional hypothesis look like? One possibility is to train a neural network to learn a mapping between words and their distributional features (e.g. words occurring in their contexts). For example, a feed-forward neural network can be used to predict a word given its co-occurrence context. The resulting representations the network learns in order to do this contain surprisingly rich semantic information (Bengio et al., 2003; Mikolov et al., 2013a; Pennington et al., 2014). The most popular of prediction-based models, a family of models often referred to as Word2Vec (Mikolov et al., 2013a), has become a popular off-the-shelf tool for learning word representations from text in machine learning applications. The representations learned by models in the Word2Vec family (e.g. Skip-gram) outperform a number of publicly available word representations in a benchmark test that includes 8869 semantic and 10675 syntactic questions (Mikolov et al., 2013a). However,

the Skip-gram model raises some concerns with regards to being taken seriously as cognitively plausible models of semantic development. For example, Word2Vec implementations contain a number of optimizations to speed training on large corpora, but some of these optimizations seem unlikely to be the way the human brain performs prediction-based learning. One requirement for training Word2Vec's Skip-gram model is knowing beforehand the frequency of words in the corpus (such that relatively frequent words can be downsampled), knowledge that is inaccessible in online learning circumstances. Another concern is Skip-gram's negative sampling procedure (Mikolov et al., 2013b), where for each prediction, only a subset of possible words are sampled from the vocabulary, including the correct next word, and others drawn from a distribution that does not include the correct word. This procedure requires knowing the correct prediction before the outcome of the prediction is computed. While this speeds training and increases performance in a machine learning context, there is no evidence for such a complex memory-based process in online human learning. A number of other optimizations (such as using the current word to "postdict" previous words in the stream) have no current basis in theories of human language processing, though this of course does not mean that such processes are impossible.

There are other neural networks that might serve as more plausible candidates for theories of semantic knowledge acquisition than Word2Vec. For example, the Simple Recurrent Network (SRN) (Elman, 1990) learns representations of words by predicting a word given a context and updating model parameters to minimize the prediction error. The first studies of the SRN showed that it could learn to predict sequences, and that doing so enables learning about the structure of the items in those sequences (Elman, 1990, 1991; Cleeremans and McClelland, 1991). For example, Elman (1991) showed that the SRN could learn the regularities of an

artificial linguistic corpus composed of thousands of sentences constructed following an extremely simplified English grammar composed of nouns, verbs, articles, and prepositions. Elman showed that the SRN could learn to predict the "correct" words in terms of following the grammatical rules and semantic constraints that were used to generate the corpus, such as noun-verb number agreement, even in cases where the verb was separated from the noun by multiple embedded clauses. Furthermore, its ability to track number agreement diminished as the length of intervening words grew larger, and this reflects experimental observations in humans.

The SRN's success at this task was due to its ability to compress sequential information into a compact distributed representation in the hidden layer. In a distributed representation, a concept is represented by a pattern of activations across an ensemble of units; by design, no single unit can convey that concept on its own. Elman showed that the similarity structure between the learned distributed representations can be interpreted as a measure of grammatical and semantic similarity between the words they represent. However, like previous researchers investigating feedforward models, Elman used an artificial and simplified corpus, and therefore left open the question of whether the SRN can scale up to noisy naturalistic language input. Recent large-scale language modeling efforts using written language corpora show that SRNs can surpass previous state-of-the-art models based on n-grams (Mikolov et al., 2014). More recently, Huebner & Willits (2018) demonstrated that an SRN can learn to predict sequences of noisy, naturalistic child-directed speech and in so doing acquire word representations that encode structured knowledge about semantic category membership. Their research shows that the principles demonstrated by Elman (1991) does not depend on the cleanliness of the artificial dataset, and that there is sufficient structure in the input that children receive to support formation of semantic categories. Huebner & Willits (2018) came to this conclusion after

conducting a large number of analyses. In what follows, I provide a brief summary of their observations, organized into two sections. Each section is associated with a different question about the nature of the representations that the SRN learned: First, how abstract is the knowledge that the SRN has acquired? Second, to what extent is the SRN's knowledge hierarchically organized?

Abstract Knowledge

An important theoretical debate in concept acquisition concerns the abstractness of knowledge. Essentially, the question is whether knowledge consists primarily (or exclusively) of a rich set of associations between sensory-motor features, or instead also consists of abstract, amodal concepts that bind those features together. Waxman and Gelman (2009) succinctly describe this as a debate between two metaphors. The first is "*child as data analyst*," whereby language acquisition occurs because of children's amazing statistical learning skills and their ability to build webs of associations of a wide variety of perceptual inputs and motor actions. This is contrasted with the "*child as theorist*" metaphor, whereby children begin with and/or build up theories about the world involving rich conceptual knowledge structures, and these knowledge structures play a critical role in structuring language acquisition. Waxman and Gelman accept a role for statistical learning, but reject an exclusively "*child as data analyst*" perspective, arguing that abstract concepts play a critical role in language acquisition and knowledge representation. Neural networks, as statistical learning algorithms, are often lumped into what Waxman and Gelman call "*child-as-data-scientist*" explanations. But most neural network models that include "hidden layers", are capable of representing abstract concepts, even if they are not the same ones Waxman and Gelman would suggest.

To demonstrate that the SRN has acquired abstract knowledge, Huebner & Willits (2018) conducted a principal components analysis of the hidden layer representations for each word in the SRN's vocabulary. The first five components were assessed by tracking which words load heavily on each component. Huebner & Willits (2018) found that the first two principal components code for high level, grammatical features that are important for predicting word order. The first principal component appeared to code for nouns, and the second principal component appeared to code for whether a word tends to appear in isolation, such as onomatopoeia and interjections. After the first two principal components, the later components began encoding semantic details. Component three was effectively coding for the activity context, specifically whether the context is "eating," compared to something more akin to "playing." Nouns and verbs relating to playing, singing, reading, watching television, and the locations where those events occur, have highly positive activations, whereas nouns and verbs relating to eating have highly negative values on this component. This is not surprising as these are likely two of the most frequent and coherent events in young children's lives, and are also orthogonal in the sense that they rarely occur together.

Hierarchical Organization

Neural network models are often criticized for not representing language or concepts in a hierarchical way that is necessary for language (Fodor and Pylyshyn, 1988; Pinker and Prince, 1988; Marcus, 1998; Gershman and Tenenbaum, 2015). But it is useful to distinguish here between what a neural network can represent, and what a neural network can *learn* to represent. Any structured, hierarchical representation can be encoded in a vector representation, and can be represented in a network's weights. Neural networks with hidden layers are, after all, universal

function approximators. Thus, there is nothing about neural networks that is incompatible with a theory that says that language must be represented as a system of discrete, hierarchically-organized symbols. The question is whether any particular neural network model can learn the correct structured representation of the language from the input.

State-of-the-art neural network models excel at mapping a sequence of words to its corresponding syntactic structure (Chen and Manning, 2014), but these models need to be supplied with the set of possible syntactic structures in order to do so, and have trouble learning those structures from the ground up. Some success has been achieved by Rogers et al. (2004) and Rogers and McClelland (2008), who showed that a feedforward neural network, learning about concepts in terms of the correlational structure of their shared features or propositional content (such as *canaries* "*are yellow*" and "*have wings*") can be used to explain the apparent hierarchical nature of concepts, and argued that hierarchical-like behavior is an emergent property of distributed representations representing the relative similarity of concepts.

To investigate to what extent the representations acquired by the SRN are hierarchically organized, Huebner & Willits (2018) conducted two analyses. First, the representational similarity of 720 nouns (referred to as probe words) which had been assigned a semantic category was computed. Huebner & Willits (2018) found that probe words that are both members of the same category tended to be more similar than probe words not belonging to the same category. Moreover, representations of probe words belonging to *related* categories (e.g. MAMMALS and BIRDS; FRUITS and VEGETABLES; MONTHS and NUMBERS) were more similar than probe words that did not belong to related categories (e.g NUMBERS and MEAT). This demonstrated that the SRN learned semantic relationships both between probe words in the *same* category, and between probe words in *related* categories. Because relatedness was found to

be, on average, higher between probe words in the same category compared to related categories, the authors concluded that the SRN learned semantic relationships at two distinct levels in a hierarchy.

Secondly, Huebner & Willits performed a hierarchical clustering analysis of probe words in the same category, to investigate the extent to which representations of words *within a category* are organized hierarchically. The results for three categories (FAMILY, KITCHEN, and SPACE) were discussed. Beginning with FAMILY, the most closely related word pairs were *grandfather* and *grandmother*, and *father* and *mother*. These words were part of a branch in the hierarchical clustering which primarily included the formal terms for family members. Another branch was identified with members such as *gran, granddad, ma, dad,* which are their informal counterparts. Similarly, two distinct branches were identified for probe words belonging to the category KITCHEN. The largest two clusters appeared to be separated according to objects used to prepare food and objects which are associated with eating. Words like *microwave* and *toaster* were found to be lumped together in a cluster that was separate from words like *teapot*, *silverware*, *napkin*. Lastly, the clustering of the category SPACE was discussed. It showed that hypernyms such as *world*, *planet*, and *star* are distinctly separated from hyponyms, such as *venus* and *mars*. In other words, the hypernym-containing cluster contains words that do not refer to any particular object in space, whereas those in its sister cluster do. This provides evidence that the model can learn to separate between concrete objects and categories containing those objects. Because organization was found at multiple levels (e.g between and within categories), Huebner & Willits (2018) concluded that the SRN was able to capture complex hierarchically organized semantic relationships between probe words.

Conclusion

In sum, the modeling results of Huebner & Willits (2018) showed that complex and highly organized semantic structure emerges automatically from learning the statistical regularities of child directed speech, supporting the idea that a neural network-like instantiation of the distributional hypothesis might explain aspects of semantic development. While answering many questions about the scalability of neural networks to large, and noisy speech input, many new questions were raised. For example, do children acquire the same categories that the SRN acquires? Does the SRN benefit from age-ordered presentation of input? The work presented here specifically addresses the latter question.

CHAPTER 2: AO-CHILDES & THE SRN

Huebner & Willits (2018) trained their SRNs on transcripts of child-directed speech that had been ordered by the age of the target child (the child spoken to). This was done to preserve as much as possible the way in which actual children experience language. While Huebner & Willits (2018) were concerned with psychological plausibility, they did not explicitly investigate whether training on age-ordered input actually influenced the representations learned by the SRN. The possibility that preservation of the age-order during training might facilitate acquisition of semantic category knowledge was raised, but the question was not followed-up. The goal of this work is to demonstrate that this is indeed the case, and to provide a detailed mechanistic explanation. While I could have chosen the corpus used by Huebner & Willits (2018) as a testing ground for such analyses, I chose not to. There are two reasons for this: First, not all transcripts used by Huebner & Willits (2018) were annotated with information about the age of the target child. To preserve the maximum amount of training data, Huebner & Willits (2018) randomly inserted transcripts into their corpus for which age information was unavailable. Because my primary interest lies in the effect that age-ordered presentation of child-directed speech might have on the model during training, I need to work with a corpus where the age-order is preserved as best as possible. This means transcripts for which no age information exists must be discarded. Secondly, the input that was used to train the SRNs in the Huebner & Willits (2018) study, while noisy and naturalistic, was a heavily processed version of the raw transcripts. The processing steps chosen by Huebner & Willits (2018) may have masked any benefit on the model that due to the age-ordered presentation of transcripts. In sum, I need a different corpus than that used by Huebner & Willits (2018).

This chapter serves three purposes: First, I describe the steps I have taken to create a corpus of child-directed speech that does not suffer the limitations of the corpus used by Huebner & Willits (2018). Second, this chapter provides a detailed description of the architecture and training regime of the SRN, and the task used by Huebner & Willits (2018) to quantitatively evaluate semantic categorization performance. Although the methods were previously described by Huebner & Willits (2018), I reproduce them here because they are the building blocks of all simulations described in this work. Third, I describe a replication study of Huebner & Willits (2018) showing that the SRN can acquire semantic categories given the novel corpus as input. While no novel questions are addressed in this chapter, a successful replication was essential in ensuring that the methodological aspects of this study were sound and that the results obtained by Huebner & Willits (2018) also hold for a novel corpus.

AO-CHILDES

As noted in chapter 1, a major criticism of previous work showing that neural networks learn abstract and highly structured knowledge is that these demonstrations have tended to use small, artificial datasets that do not capture the real noise and complexity of speech to children. To address this problem, Huebner & Willits trained the SRN on the American English section of the CHILDES database, a collection of transcripts of interactions with children in various situations (MacWhinney, 2000). The CHILDES database contains a mixture of transcriptions of structured in-lab activities (such as book-reading, mealtime, and playing with toys), free play in the lab, and in-home recordings. While noisy and naturalistic, Huebner & Willits (2018) performed additional processing of the raw transcripts that may have influenced their results. Specifically, differently spelled forms of the same word were converted to the same form; plural,

possessive and diminutive morphemes were split from nouns; plural past-tense and ongoing morphemes were split from verbs; split morphemes were left in the corpus as separate units; finally, proper names were replaced with symbols signifying the gender of the person in question.

In the replication study, I followed the same procedure as Huebner & Willits, except that I did not regularize spelling, perform any morphological parsing, or replace proper nouns with gender-informative symbols. To create the new corpus, I first obtained all transcripts in the CHILDES database[1] that involve children 0 to 6 years of age from American English speaking households and excluded those for which no age information was available[2]. After removal of non-adult speech, I obtained 3,251 transcripts containing 22,448 types, and 5,113,856 tokens. One hundred randomly chosen transcripts, containing 64,007 word tokens, were set aside during training to assess the SRN's ability to predict word sequences not encountered during training. Considering that a typical working-class American child receives approximately 6.5 million words per year (Hart and Risley, 2003), the training corpus represents approximately 4–10% of the amount of lexical input of a 3-year-old child (there are large individual differences largely predictable by socio-economic status). The documents of the corpus were organized by the age of the child spoken to, such that each model experienced the input in an age-appropriate way, receiving the input a 6-month-old hears, then a 7-month-old, then an 8-month-old, etc.

The transcribed corpus was tokenized (split on spaces) with sentence-boundary punctuation (periods, exclamation marks, commas, and question marks) left in the corpus. This was intended to serve as a very crude way for representing the pauses and prosody that tend to

[1] retrieved from childes-db.stanford.edu on Dec 1, 2017

[2] Huebner & Willits (2018) did not exclude transcripts for which no age information was present, and therefore worked with a slightly larger corpus.

accompany utterance boundaries. Contrary to Huebner & Willits (2018), I performed no further processing, to leave intact as many naturalistic properties of the corpus as possible. For simplicity I will refer to the resulting corpus as AO-CHILDES to indicate that the transcripts it contains are ordered by the age of the target child (AO is short for age-ordered). The total size of AO-CHILDES is approximately 200,000 words smaller than the corpus used by Huebner & Willits. This was primarily due to the exclusion of transcripts for which no age information was available. The age-distribution of transcripts in AO-CHILDES is shown in figure 2.1.

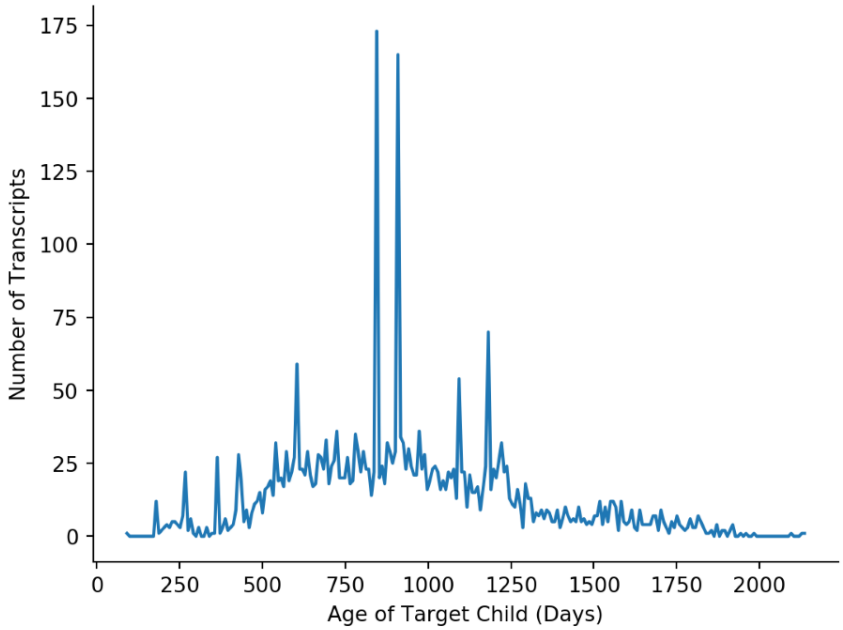

Figure 2.1
Histogram relating age of the target child to the number of transcripts in AO-CHILDES.

The CHILDES database is not perfect as a representative sample of the full range of activities that parents participate in with their children or the variety of language used during those activities, but is instead a useful approximation. Indeed, the relatively constrained set of

activities that occur in CHILDES ought to hinder learning of useful semantic structure, and thus make positive results all the more impressive.

Vocabulary, Probe Words and Categories

To reduce training time and simulate the fact that children are unlikely to know the lexical form of the lowest frequency items in the corpus, Huebner & Willits (2018) limited the model's vocabulary to the 4,096 most frequent word types. I followed the same procedure. All words that were not included in the vocabulary were replaced with the symbol 'UNKNOWN' before being input to the model. Given that word distribution obeys a power law, this only affects less than 0.8 % of all total word tokens in AO-CHILDES. Huebner & Willits noted that the size of the vocabulary does not significantly alter their results, noting that a vocabulary size of 12,511 (each word occurs at least twice) did not alter learning outcome. In AO-CHILDES, each word occurs at least 28 times.

In order to address the question of whether the SRN has learned abstract and structured knowledge, Huebner & Willits (2018) chose to investigate the model's knowledge of a set of probe words belonging to a set of pre-identified categories. The set of probe words were chosen from the vocabulary by, (1) choosing the subset of word forms which could be nouns (even if, in practice they appear more often in verb form, such as *jump*), (2) choosing the subset of those that refer to a concrete object, and (3) choosing the subset of those that unambiguously belong to a semantic category from which at least six other words belong, according to a set of human raters. For example, *apple*, *orange*, and *banana* (along with many other fruit words) were included because they belonged to a large category of items that contained at least six items. Because of the smaller size of AO-CHILDES compared to Huebner & Willits (2018), and the lack of

additional post-processing steps, I obtained only 532 words belonging to 28 categories, which is 188 fewer words than those obtained by Huebner & Willits (2018). Following Huebner & Willits (2018), I will refer to these words as probe words to differentiate them from the words in the vocabulary that were not used during evaluation of semantic categorization performance. A complete list of probe words and their categories is available in Appendix A.

The Simple Recurrent Network Architecture

The Simple Recurrent Network (SRN) is an artificial neural network that contains an input, a hidden, and an output layer, in addition to copy connections linking the hidden layer to the input layer at the next time step (Elman, 1990). The hidden layer learns distributed internal representations of the input, and the recurrent connectivity allows these representations to encode information from previous time steps. This means that the activations at the hidden layer are not a simple representation of the input stimulus, but rather the input stimulus in the context in which it occurred.

A schematic of the SRN's architecture is shown in figure 2.2. For each time step, the SRN received as input a localist representation of a single word drawn sequentially from the training corpus. The localist representational scheme ensures the model has no access to information about word similarity (phonological, semantic, etc.) at this stage. This is done by filling the input vector with zeros at every of 4,096 positions (corresponding to the vocabulary size) except for the position uniquely assigned to the current input word. The goal of this scheme is not to claim that children do not utilize additional sources of information about input words, but to test just how rich a child's semantic knowledge could become based on lexical distributional information alone. This localist representation of a word is used to index a

distributed representation of size 512, before being fed to the hidden layer. This step is strictly not necessary, but reduces computational overhead without quantitatively or qualitatively affected the learning procedure.

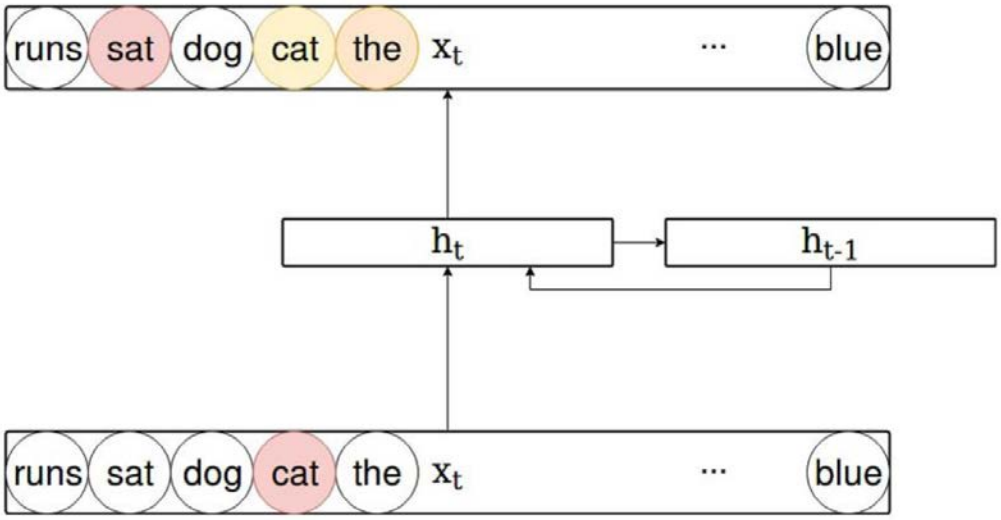

Figure 2.2
The SRN architecture is conventionally depicted as consisting of three layers. In our work, the one-hot input vector (shown in the bottom of the figure) was used to index into a trainable word embedding layer before being fed to the hidden layer. This converts the localist representation of a word to a distributed representation. Colors depict the level of activation from 0 to 1 (yellow to red).

The activations at the hidden layer (512 units) are the result of multiplying each input unit's activation by the weighted connections from that input unit to each hidden unit. Critically, the hidden units activations at the previous time step are added, weighted by the recurrent connections. Lastly, each hidden unit activation is transformed by the hyperbolic tangent non-linearity to constrain its activation between -1 and +1[3]. The hidden layer activations are then sent via a third set of weighted connections to the output layer (again containing 4096 units). The net input into each output unit is first transformed by exponentiating e to the weighted input

[3] The results obtained with a sigmoid activation function are nearly identical.

(effectively flattening the distribution. of activations across the output units). These output activations are then transformed, via the softmax operation, into a posterior probability distribution representing the SRN's predictions about which word should come next.

Training Regime

Contrary to conventional training methodology in which input sequences are presented to the model from multiple iterations over the whole corpus, Huebner & Willits (2018) presented input sequences to the SRN from multiple iterations over small partitions of an age-ordered corpus. Specifically, the age-ordered corpus was split into 256 equally sized partition such that the number of words in each partition roughly corresponded to the number of words heard by children in 1 day (Hart and Risley, 2003). Huebner & Willits (2018) then trained their SRNs on each partition in order. For example, partition 1 was seen 20 times, and then partition 2 was seen 20 times, and so on. This was done to lend cognitive plausibility to their training regime. Huebner & Willits (2018) reasoned that it is more likely that children consolidate linguistic experiences across time periods spanning hours or days rather than months or years. Because I am more concerned with the *order* in which partitions are trained on, rather than cognitive plausibility, I split AO-CHILDES into 2 equal sized partitions instead. This will simplify interpretation of experiments described in subsequent chapters in which partition order is explicitly manipulated. Because the split occurs at the midpoint, partition 1 contains speech to younger children (1-3 years of age), and partition contains speech to older children (3-6 years of age). During training, the SRN was exposed to partition 1 20 times, before being exposed to partition 2, also 20 times. Because the number of words in each partition is considerably greater

than the number of words a child hears in a day, this training regime can no longer be viewed as simulating memory consolidation of linguistic experiences operating on a *daily* schedule.

Following Huebner & Willits (2018), I trained several (n=8) SRNs with different random seeds during weight initialization. Weights were initialized with a truncated normal distribution with mean zero and standard deviation $1 / \sqrt{m}$, where m is the number of units in the layer above. A bias unit was used at the output layer and its weights were initialized to zero. For every word in the training corpus, I feed into the model a sequence consisting of the word and the six words immediately left of it. The input was fed through the model (as described above) and resulted in a probability distribution of predictions for the next word in the sequence. I used the cross-entropy operation to compare a model's predictions to the correct answer, which is equivalent to the negative log of the probability assigned by a model to the correct answer (i.e., the next word in the sequence). I used truncated backpropagation through time (Werbos, 1990; Williams and Peng, 1990) to compute the partial derivative of each layer's activations with respect to the weights, and used these to update the weights in the direction that minimized prediction error. This procedure was followed sequentially for each input sequence. I set the learning rate to 0.01 and used Adagrad optimization (Duchi et al., 2011) to adapt the learning rate so that infrequently changed weights receive a greater update than those changed more frequently. Weights were adjusted using mini-batch training, in which weight updates only occurred after the accumulation of prediction errors from 64 sequences. In this way, the weight update reflects the average prediction error computed for all 64 sequences in the mini-batch. While the primary motivation for using mini-batching is to speed model training, the cognitive and neural plausibility of mini-batch learning is contestable. To address these concerns, Huebner & Willits (2018) tested a range of different mini-batch sizes, and found that sizes greater than 64 led to slightly worse results,

with no noticeable differences (other than in training time) for smaller mini-batch sizes, including a size of 1. Thus, this detail in the model speeds training without a cost in terms of a qualitative or quantitative change in the model's behavior and thus calling into question its cognitive or neural plausibility.

Computing word representations

Typically the representation of a word in the SRN is understood as the pattern of hidden layer activations that results when a word is fed into the model. However, Huebner & Willits (2018) opted for a slightly different approach, taking advantage of the fact that the SRN can represent sequences, consisting of a word and its context. Because the SRN never sees a word in isolation, Huebner & Willits (2018) considered that simply retrieving the activations at the hidden layer given a word at the input layer in the absence of any context, might distort the knowledge that the SRN has actually learned. Instead, the representations were constructed in such a way that they preserved the structure of the input that the model has actually encountered during training. To compute a single word's representation, all sequences in the corpus in which the word occurs in the last position during training were re-input to the SRN, and the resulting pattern of activations at the hidden layer were saved. Huebner & Willits (2018) defined a word's representation as the average of those hidden layer activation vectors. In all subsequent analyses, I use the term 'word representation' to refer to these vectors. Offline analyses showed that representations of probe words computed in this way contain more information about semantic category membership than representations that are simply a word's pattern of activation at the hidden layer.

Replication of the Quantitative Analysis in Huebner & Willits (2018)

While Huebner & Willits (2018) performed a number of qualitative analyses to demonstrate that the knowledge the SRN has acquired is both abstract and hierarchically structured, I restrict evaluation of the SRN to the quantitative analysis employed by Huebner & Willits (2018). My aim is not to investigate whether utilization of a novel corpus, AO-CHILDES, influences the abstractness or the hierarchical organization of the learned representations, but to confirm that semantic category acquisition still occurs when AO-CHILDES is used as the input. After all, AO-CHILDES is a smaller and less processed version of the input used by Huebner & Willits (2018), and may return unexpected results.

The quantitative analysis has been adapted from Huebner & Willits (2018) without modification[4]. Because it is my primary method for evaluating semantic categorization performance, and because I have made frequent use of it in the experiments described in subsequent chapters, a detailed description follows. The analysis is best understood as a semantic classification task in which two probe words are judged to be in the same category. Judgments are based on a 532 by 532 matrix, **S**, of the similarity of all probe words with one another. In this task, all word pairs' similarity scores were compared against a decision threshold and used to guess if the two words belong to the same semantic category. I analyzed these results in a signal detection framework, computing hits, misses, correct rejections, and false alarms for each probe-word pair at multiple similarity thresholds (r, between 0.0 and 1.0 with step size 0.001). In other words, if two probe-words, represented by the row index i and col index j, belong to the same category, and $\boldsymbol{S}_{i,j} > r$, a hit is recorded, whereas if $\boldsymbol{S}_{i,j} < r$, a miss is recorded. On the other hand, if the two probe-words do not belong to the same category, either a correct rejection or false

[4] However, the number of probe words used here is smaller, because the input (AO-CHILDES) is smaller, and the vocabulary is not identical to that used by Huebner & Willits (2018)

alarm is recorded, depending on whether $\boldsymbol{S}_{i,j} < r$ or $\boldsymbol{S}_{i,j} > r$. For each probe word, I calculated the sensitivity and specificity, and averaged the two to produce the balanced accuracy. This procedure eliminates bias resulting from the fact that the vast majority of probe word pairs do not belong to the same category. The measure of interest was the average of all the probe-words' balanced accuracies at the similarity threshold which yielded the highest value.

I calculated the balanced accuracy at equally spaced intervals during the training for each of the 8 SRNs. I averaged the resulting trajectories and plotted them in figure 2.3. Initial performance is close to chance[5] and increases as a function of training time, as expected. The average end-of-training balanced accuracy is 0.73 ± 0.002 (mean $\pm$ standard deviation), which is 3 points higher than the mean (0.70) reported by Huebner & Willits (2018). This means that their analysis has been successfully replicated here, using a less post-processed corpus, AO-CHILDES, and under a small modification of the training regime (2 vs. 256 partitions). What about the greater performance at the end of training? It is best not to directly compare the end-of-training balanced accuracy, because only a subselection of the probes used by Huebner & Willits (2018) were used here. While the task is identical, the test items are not, and this makes comparison impossible. In fact, I expected performance to be slightly larger for the following two reasons: First, because of the lack of corpus post-processing steps used here, the most frequent 4,096 words used to make up the vocabulary, are no longer identical to those used by Huebner & Willits (2018). Because of this shift in the vocabulary, some probe words were excluded. Because the excluded probe words are the least frequent of the full set of probe words,

[5] Chance performance is 0.5. However, the balanced accuracy actually obtained before training is approximately 0.6. This occurs because representations are for *sequences* in which probe words occur (in the last position) rather than probe words in isolation. Above-chance performance at the beginning of training is not due to any knowledge encoded in the weights, but due to information present in sequences used to obtain probe word representations.

their exclusion likely resulted in a modest improvement in performance. Second, iterating 20 times over 2, rather than 256 partitions enables the SRN to integrate information across larger chunks of the corpus. Offline analyses confirmed that both factors independently raise end-of-training performance. Because these factors have been shown to influence performance, little can be said about the effect that the skipping of the corpus post-processing steps performed by Huebner & Willits (2018) may have had. While it is tempting to conclude that performance was not negatively impacted, it is more likely the case that the performance improvement provided by the exclusion of infrequent probe words and iterating over larger partitions masked any negative influence on performance due to the skipping of corpus post-processing steps. Nonetheless, the results obtained here at least suggest that semantic categorization was not severely limited by the absence of additional corpus post-processing.

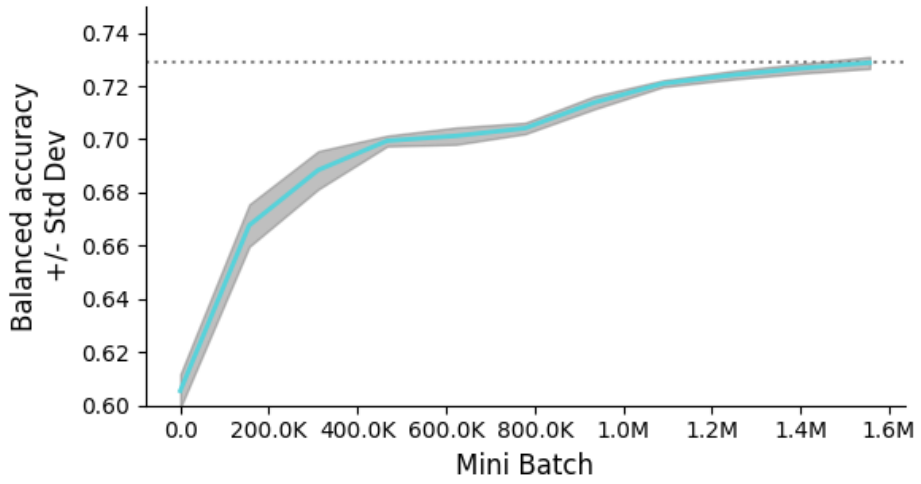

Figure 2.3
Average balanced accuracy as a function of training time (in minibatches) of 8 SRNs trained in age-order on 2 partitions of AO-CHILDES.

It is important to note that the categories underlying the semantic classification task were chosen by Huebner & Willtis (2018) and revised based on the explicit judgments of adult experimental participants. A breakdown of the balanced accuracy by category showed that performance varied greatly with the category. Huebner & Willits (2018) cautioned that the categories with lower scores, while quite real to adults, may be less important to children and thus not frequent or consistent in child-directed speech. Though, it is possible that these categories are less "real" in either a natural or psychological sense, and that these lower scores therefore reflect exactly how the model should perform. Follow-up corpus analyses and behavioral experiments (with children and adults) can further investigate the natural or psychological reality of these categories, and assess the extent to which different models predictions about the cohesiveness of a category reflect the representations that children acquire.

Conclusion

The purpose of this chapter is to describe and validate the methods that will be used in the experiments reported in subsequent chapters. First, I recorded the creation of AO-CHILDES, a corpus that I will use to investigate questions about the influence of age-ordered presentation of training examples. Additionally, I detailed the architecture of the SRN, the training regime, and the task used to evaluate semantic categorization performance. Each will be frequently used in the experiments described in this work. Lastly, I showed that SRN training on AO-CHILDES results in end-of-training performance that is in the same range as those obtained by Huebner & Willits (2018). I conclude that AO-CHILDES is a good corpus to further investigate semantic category acquisition. Due to the lack of post-processing and insertion of transcript for which no

age information is available, AO-CHILDES is the ideal testing ground for questions about age-related changes in the linguistic input which may affect semantic categorization performance.

CHAPTER 3: THE AGE-ORDER EFFECT

In this chapter, I demonstrate that training on AO-CHILDES in order (partition 1 first, partition 2 last) increases end-of-training semantic categorization performance, compared to training on AO-CHILDES in reverse order (partition 2 first, partition 1 last). Before discussing the results, I explain the motivation behind this study, which includes a discussion of child-directed speech, and a brief overview of related observations in the cognitive modeling literature. Underlying the question of how the learning trajectory is influenced by the order of the input are two distinct questions: How does the input change over time? Second, how is a neural network model influenced by the order of the training data? I will explore these two questions in turn.

Child-Directed Speech

Because AO-CHILDES is largely made up of child-directed speech, it is worth discussing how child-directed speech (CDS) differs from adult speech. There are numerous differences, such as larger pitch contours, lengthened vowels (Fernald & Kuhl, 1987; Gleitman, 1984) and restricted use of complex constructions (Broen, 1972; Furrow, Nelson, & Benedict, 1979; Newport, Gleitman, & Gleitman, 1977; see Pine, 1994; Richards, 1994 for reviews). Individuals engaging in CDS, are also more likely to restrict the range of conversational topics, choice of words and grammatical abstractions (Snow & Ferguson, 1977; Lieven, 1994), and make longer pauses between utterance boundaries (Gallaway & Richards, 1994). While numerous benefits of CDS have been found on early language acquisition (Golinkoff and Alioto, 1995), others have shown that children appear to learn language just as well when their primary caregivers do not use CDS (Schieffelin and Ochs, 1986). It is quite remarkable how little

variation in the input seems to affect learning outcomes. The question of what role the input plays in language acquisition is hotly debated. Clearly some input is needed, but how much and how do qualitative differences like type/token ratio affect learning? There are numerous studies that highlight the role that individual variation in language exposure has on vocabulary growth (e.g. Huttenlocher et al. 1991; Hart & Risley, 1995). Findings like these suggest that the quality of the language input can have a strong influence on learning outcomes. This is one of the many reasons I have become interested in the role that CDS might play in language acquisition.

As children mature in their language abilities, CDS is gradually replaced by adult speech. The speech that a newborn hears is an extremely skewed sample of the speech that a 3 year old might hear, and an even more skewed sample of the speech a 6 year old might hear. How exactly is CDS gradually replaced by adult speech? Do speakers track the age of the children they speak to, or do they track their linguistic abilities, or both? This is an important question, because over-extending the time in which linguistic input is simplified may actually delay acquisition. For example, learners provided exclusively with simplified speech may never learn more complicated constructions in their language. On the other hand, providing more advanced examples of language from the beginning may delay learning due to the large number of hypotheses that would have to be considered (Cameron-Faulkner et al., 2003). This presents a conundrum to the language teacher: How should samples from a language be chosen to maximize learning outcomes? Several computational studies have approached this question. A brief review follows below.

Starting Small

The idea that a learning system can be guided to a solution by selectively modifying the order in which learning experiences are presented can be traced back to work by Elman (1993). Using a simple recurrent network to predict tokens in a small hierarchically structured artificial language, Elman found that learning took place only when either the input or the model was constrained such that initially only simple linguistic dependencies can be learned. In the latter case, the model was restricted by a limited memory span at the start of training, and with more training this constraint was gradually relaxed. He coined the expression “starting small” to refer to this strategy. Elman concluded that “starting small” might allow human brains to learn patterns from their linguistic experiences which otherwise might be unlearnable. Starting with an initial simpler state, human brains might be predisposed to learn only those abstractions which they are capable of learning. In part, his claim was targeted directly at nativists: The role that domain-specific knowledge is supposed to play may be attributed instead to the initially undifferentiated and limited cognitive abilities of young infants. Even if this assertion is not born out, having shown that the same benefits can be provided by staging the input in order of increasing linguistic complexity, Elman clearly demonstrated the importance of “starting small".

In a follow-up examination using the same model and a larger range of artificial language corpora as input, Rohde & Plaut (1999) showed that “starting small” does not always facilitate learning. In fact, the majority of their simulations showed that initially restricting the grammatical complexity of the input actually reduced learning. The severity of the learning impairment was most severe when the input was made more naturalistic via addition of semantic dependencies. Only in extreme cases in which the input was devoid of such semantic constraints, did “starting small” provide an advantage. Rohde and Plaut concluded that “starting small” may

be of little consequence for learning the grammatical structure of a human language, due to the large number of semantic dependencies present in naturalistic languages. Furthermore, they speculated that simple recurrent networks naturally start with a limited memory, and that it is naturally expanded over the course of training. Further simplifying the input should have little effect and might disrupt the network's natural tendency to "start small". In Rohde and Plaut's view, the basic notion behind "starting small" is useful, although in a different form than Elman imagined: Computational models, and human brains, prior to having experienced any input, may be sufficiently unorganized to implement "starting small" out-of-the-box. In their account, the benefit of staging the input has to do with minimizing interference between novel complex experiences and established memories of simpler experiences. If the learning apparatus is pre-equipped with a filter for complex input, then there is no need for additional restrictions in the input. In fact, such restrictions might even be harmful. Lastly, Rohde and Plaut note that simplified input may actually be beneficial for learning the meanings of utterances, and that there likely exists a tradeoff between learning the grammar and the association between meanings and surface forms of a language. Clearly, the issue is much more complex as initially proposed by Elman.

Since the publication of Rohde & Plaut (1999), the issue has remained largely untouched. It is likely that the computational resources at the time did not permit a more thorough investigation of "starting small" in large corpora of naturalistic language. Nonetheless, with the arrival of such resources and the number "deep learning" breakthroughs training statistical models on large text corpora, few if any researchers have revisited the issue. Currently, researchers pursuing the promise of data-driven language acquisition disregard altogether the temporal organization of the input by shuffling the order of the input with every training epoch.

The rationale behind doing so is to reduce the risk of overfitting. While overfitting is an important concern for "deep learning" practitioners, and shuffling the input often eliminates unwanted statistical irregularities, a smarter approach to reordering their data is not frequently considered. However, I identified several notable exceptions. I discuss each below.

First, in his doctoral thesis, completed after the work of Rohde & Plaut (1999), Rohde notes that training using staged input resulted in "better learning" compared to using the full input. The model under investigation was similar to the SRN used previously except that a different artificial language was used to train it. Little more was said about his findings, most likely because training required several months to complete and retraining using different input conditions would have been prohibitively time-consuming. He noted however that such a result was in accord with Rohde & Plaut (1999) who claimed that "starting small" should facilitate learning of the *meanings* of words, phrases and longer utterances in the language.

Bengio et al. (2009) compared training on the full input with no apparent order to a curriculum learning strategy whereby input is separated into stages according to some measure of task difficulty. For example, a neural network trained to categorize 2D shapes benefited from a curriculum strategy whereby the first half of training examples were sampled from a training distribution with less variability in shape than the full training distribution. Similarly, a neural language model trained to categorize word sequences as grammatical, given samples of a target language, achieved a lower test error when trained incrementally on sections of the input ordered by the size of the vocabulary. The authors suggested that a curriculum learning strategy acts both as a way to find better local minima and as a regularizer (performance improvement was evident primarily on the test data with little improvement on the training data). Moreover, curriculum

learning should speed training because less time is wasted with noisy or harder to predict examples.

Mikolov, in his doctoral thesis (2012), discovered that ordering chunks of several standard written text corpora based on the perplexity obtained by a 2-gram model, trained on the same data, and excluding any chunks with very large perplexity, improved learning considerably. This kind of manipulation of training data is roughly equivalent to sorting the data by the uncertainty associated with word prediction. Explaining his motivation for the experiment, Mikolov noted that complex patterns in the data are based on simpler patterns and that "these simple patterns need to be learned before complex patterns can be learned". This is in line with the "representational trajectory hypothesis" proposed by Clark (1993). Reviewing Elman's work on "starting small", Clark suggested that restricting the early input to simple examples might direct a learner's representational device into a direction more suitable for learning more complex linguistic abstractions. Specifically, Clark asserted that failure to learn lower order regularities in the training data should prevent learning of higher order features that depend on those lower order features.

Lastly, Graves et al. (2017) studied the usefulness of various learning progress signals for selecting the next task to train on, and found that some signals can lead to significant gains in curriculum learning efficiency compared to a uniform sampling approach. Prediction gain, which selects the next sample based on the change in loss before and after training on a sample, was found to perform best for maximum likelihood training. This measure is used as an indicator of learning progress and is therefore calculated during training. In this way the curriculum is updated in an online fashion, based on some measure of learning progress, rather than precomputed. The researchers also noted that the uniform sampling approach presents a strong

baseline, and that this may be due to an implicit curriculum inherent in gradient descent training. Because learning is dominated by gradients from tasks which are learned fastest, the direction of gradient descent tends to be in the direction of where the most progress can be made. As such, automating a curriculum may be best viewed as utilizing a learner's natural tendency to learn those tasks which at any given time during training result in the largest training progress.

The current study

I trained two groups of SRNs in either age-ordered or reverse age-ordered training condition. SRNs in the age-ordered training condition were exposed 20 times to each training example in partition 1, and were then exposed to 20 times to each training example in partition 1. In the reverse age-ordered training condition, training examples from partition 2 were seen first, followed by training examples from partition 1. The total amount of tokens trained on in each condition is identical; the only difference is in the order of presentation. Because speech to younger children (partition 1) tends to be less complex than speech to older children (partition 2), age-ordered training on AO-CHILDES can be viewed as 'starting small'. Compared to Elman (1993), however, this work focuses on how 'starting small' might affect semantic category, rather than syntax acquisition. In fact, it is impossible to study syntax acquisition without an artificial grammar because the generative model that underlies the child-directed speech in AO-CHILDES is unknown. While I will compare the ability of the SRNs on the average per-word perplexity, it must be kept in mind that this measure is strictly an indicator of a model's fit to the data, and not to the underlying syntax. In other words, the perplexity represents a measure of fit to the training data (or test data), rather than knowledge of the abstract principles of English syntax. This is an important tradeoff that must be considered when using naturalistic speech to

train the SRN. While using AO-CHILDES hampers my ability to evaluate acquisition of English syntax, it enables me to ask how the input that children *actually receive* influences semantic category learning in the SRN. Constructing such a corpus artificially would create many concerns about whether 'starting small' could actually apply to children learning in the real word.

An important innovation in this work is that both next-word prediction and semantic and syntactic category learning are evaluated in the same model. Children simultaneously learn both the structure of language and categories inherent in their input. Whether knowledge about categories and sequential structure are the outcome of a single learning process is unknown, but because formation of categories are a requirement for making useful next-word predictions, it is sensible to study the interplay between the two. Proposing that language acquisition involves learning the statistical relationships between words, requires an equal commitment to the existence of categories of words with similar contexts. These categories may come to resemble syntactic classes (e.g. nouns, verbs, adjectives, etc.) which exist because they predict similar temporal relationships, or semantic categories which exist because they predict semantic relationships. The distributional hypothesis makes no distinction between semantically and syntactically similar words.

Sequential structure prediction

First, I evaluated SRNs on their primary task, sequence prediction. To do so, I tracked the average per-word perplexity on both the training and withheld test data at several time points during training. To simplify comparison, I averaged the trajectories across models in the same condition. The results are shown in figure 3.1. In both conditions, the average perplexity drops

for both the training and test data. This demonstrates that sequence learning has occured for SRNs in both conditions. Comparing end-of-training performance, I found no difference between age-ordered training compared to reverse age-ordered training on perplexity computed on the training data (43.5 $\pm$0.01 vs. 43.5$\pm$0.11, respectively). However, age-ordered training resulted in a statistically significant ($t = 13.8$, $p < 0.0001$) improvement in end-of-training perplexity on the test data (49.6 $\pm$0.02 vs. 49.1$\pm$0.10, respectively). The difference is not big, but provides initial evidence that age-ordered presentation can influence learning outcomes. At the end of training, the models trained in age-order reached local minima that generalize better to unseen data. This indicates that age-ordering can have a regularization-like effect.

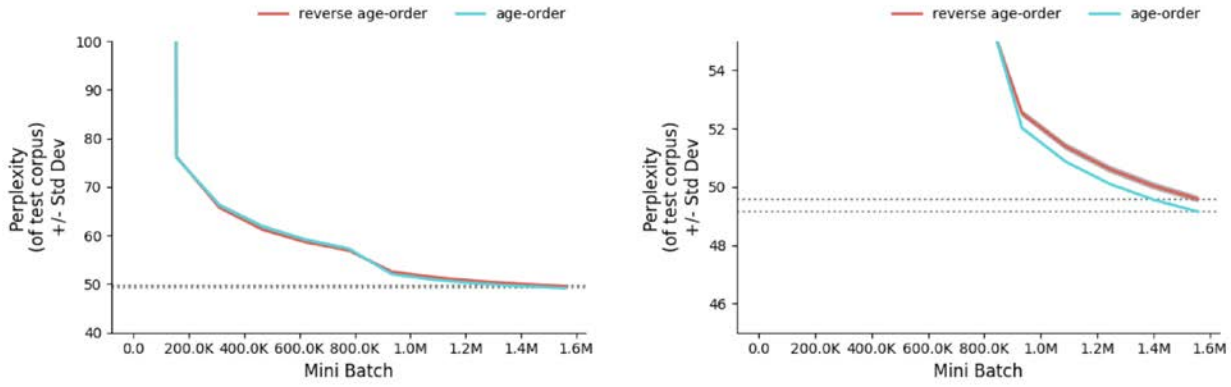

Figure 3.1
Average perplexity computed on withheld test corpus as a function of training time for SRNs trained in age-order (blue) and reverse age-order (red).

Semantic Categorization

Next, I compared the balanced accuracy of the models in the age-ordered training condition to those in the reverse age-ordered training condition. As mentioned in chapter 2, this score reflects the ability of the model to cluster probe words belonging to the same category. The model's judgements are compared to the ground truth, which was obtained via adult categorization judgments. I did not report a similar measure of clustering performance, the F1-

score, because it suffers from issues related to interpretability and produced similar results in our simulations. I plotted the results in figure 3.2. Shown in blue, the average balanced accuracy of the SRNs trained in age-order is consistently higher than the average balanced accuracy of the SRNs trained in reverse age-order. Curiously, end-of-training performance is still larger for the SRNs trained in age-order despite having been exposed to exactly the same training examples. A two-tailed t-test confirms that this difference is statistically significant, $t=7.3$ ($p<0.0001$). For simplicity, I will refer to the persistent improvement in performance exhibited by the models trained in age-order as the age-order effect. What is most interesting about the age-order effect is not its *magnitude* (it is relatively small), but that it occurred at all. Trying to answer why it did, is the subject of the remainder of this work.

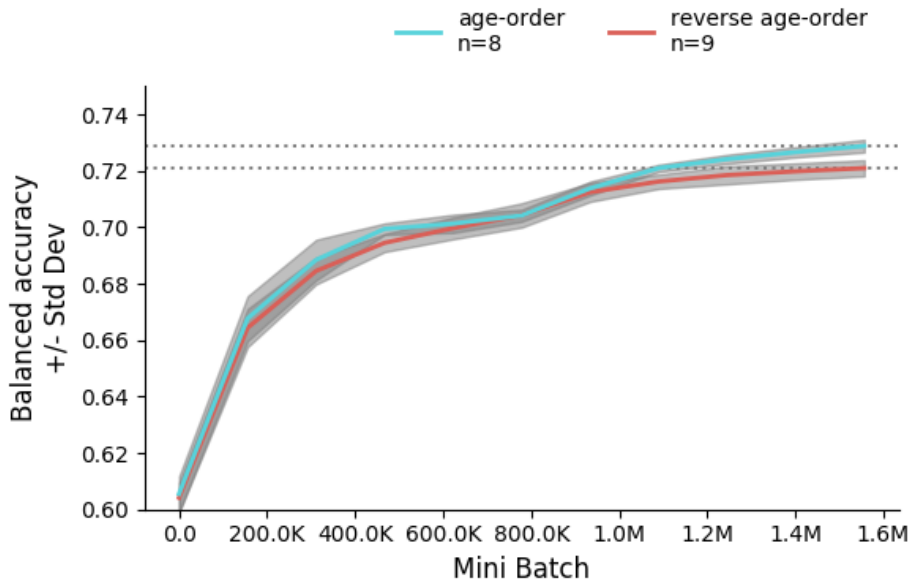

Figure 3.2
Average balanced accuracy as a function of training time (in minibatches) of 8 SRNs trained in age-order (blue line) and 9 SRNs trained in reverse age-order (red line) on 2 partitions of AO-CHILDES.

To verify that the age-order effect is actually due to the difference in the two training conditions, I re-ran the same simulation, but with the age-structure of AO-CHILDES removed. To do so, I randomly assigned half of the 3,156 transcripts in AO-CHILDES to partition 1 and the remaining half to partition 2. The training regime was kept the same; each SRN iterated 20 times over each partition in order or in reverse order. Because there is no systematic difference between the two partitions, no performance difference should be observed. The results of this simulation are shown in figure 3.3. Indeed, the two average balanced accuracy trajectories overlap for the entire duration of training. The difference in performance at the end is not significant, t=1.57 (p=0.137).

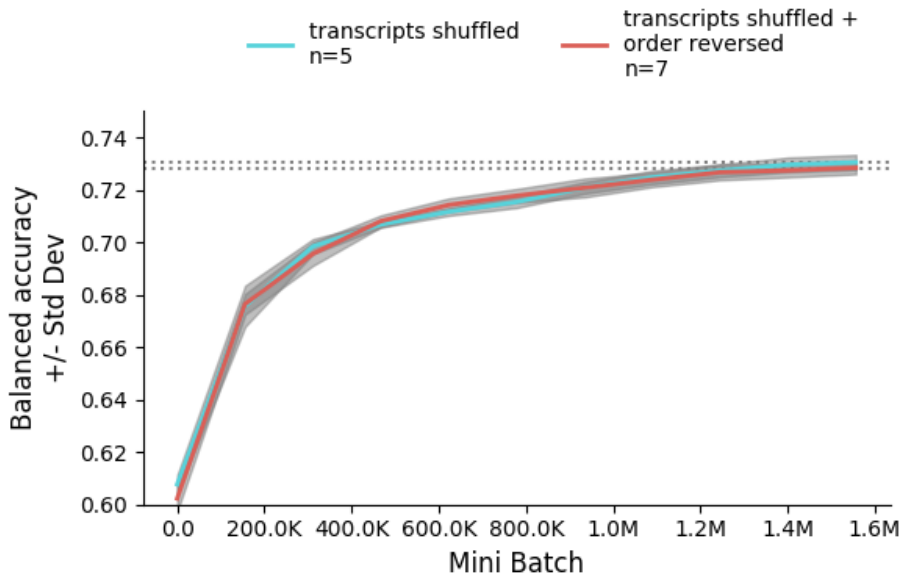

Figure 3.3
Average balanced accuracy as a function of training time (in minibatches) of 5 SRNs trained on an age-shuffled version of AO-CHILDES (blue line) and 9 SRNs trained in reverse order (red line) on the same shuffled corpus. The order of the transcripts before splitting the corpus into 2 partitions was shuffled. No performance difference was expected in this control experiment.

Drivers of the Age-order Effect

Next, I asked whether the age-order effect is driven by a large improvement in balanced accuracy for members of only a small number of categories. If so, this would suggest that age-ordering facilitates acquisition of a specific set of categories, rather than providing a universal advantage. If I found that only a small number of categories were driving the effect, I would start with a careful inspection of how the affected categories differ in their distributional properties from the rest. If, on the other hand, the advantage of age-ordered training appears to be universal across categories, I would search for an explanation elsewhere; in that case, it would be more appropriate to think more broadly about the characteristics of the input and how they interact with the learning dynamics of the model. Inspecting the balanced accuracy broken down by category, I found that there was no small set of categories that is driving the effect. Instead, I found that roughly half of the categories benefitted from age-ordered training approximately equally. Moreover, these categories tended to be categories for which the balanced accuracy was higher than the average balanced accuracy computed across categories. This suggests that age-ordering is not due to some special property of a small number of categories, but influences semantic category learning *as a whole*.

Given the improved performance of the SRNs trained in age-order, it is clear that the SRNs trained in age-order are capturing additional information that the SRNs trained in reverse age-order are not. To better understand the age-order effect, it would be useful to know whether this additional information is specific to a particular partition. To do so, I split the 532 probe words into two equally sized sets; one contains probe words which tend to occur more frequently in partition 1 (semantic-early) and probe words which tend to occur more frequently in partition 2 (semantic-late). I trained the same number of SRNs in each training condition again, and this

time tracked the balanced accuracy associated with probe words in each set separately. At the end of training, I averaged the balanced accuracies across models in the same training condition. The results of this analysis are shown in figure 3.4. The results show that the end-of-training performance advantage of the models trained in age-order (shown in blue) is largely restricted to probe words that tend to occur more frequently in partition 2 compared to partition 1. One interpretation of this finding is that pre-training on partition 1 facilitates acquisition of semantic category information present in partition 2 relative to training on partition2 'from scratch'. Without the knowledge gained during training on partition 1, the SRN appears to experience greater difficulty in extracting semantic category information in partition 2. Though, one must be careful with such an interpretation, because it is not clear whether category membership of late-occurring probes is more difficult to acquire because of some difference in the two *partitions*, or some difference between the distributional properties of late and early occurring *probe words*.

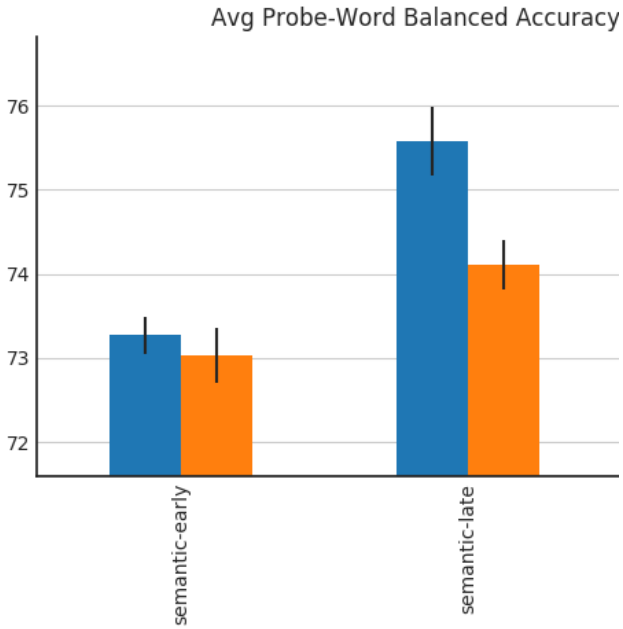

Figure 3.4
Average balanced accuracy at the end of training computed across models trained in age-order (blue) and reverse age-order (orange) for probes that tend to occur more frequently in partition 1 (semantic-early) and probes that tend to occur more frequently in partition2 9semantic-late).

Syntactic Categorization Performance

I was curious whether a similar effect of training order would influence categorization performance of words into syntactic categories. In this analysis, 830 probe words were used, which were selected based on whether they can be unambiguously classified into one of eight syntactic categories (adjective, preposition, adverb, determiner, interjection, noun, pronoun, verb). Underlying this analysis is the question of whether the age-order effect is specific to semantic categorization or whether it is related to some overall ability of the model to acquire *both semantic and syntactic* categories. The setup was identical to that in the previous simulations, except that balanced accuracy was computed on a different set of probe words and syntactic, not semantic, categories were used to judge the correctness of similarity judgements. the results are shown in figure 3.5. The models trained in reverse age-order have a strong advantage over models trained in age-order during the first half of training, when models trained in reverse age-order are training on partition 2 and models trained in age-order are training on partition 1. This indicates that syntactic categorization is greater when training on speech to older children compared to training on speech to younger children. This is not surprising, given that speech to younger children (as we will see in detail in the next chapter) is syntactically more restricted. Specifically, members of noun-noun syntactic categories occur more frequently in partition 1 than partition 2. This means that examples of conjunctions, prepositions, adjectives etc. are all more frequent in partition 2, at the expense of the number of nouns. Interestingly, this early advantage is eliminated after the models trained in reverse age-order cross the partition boundary to start training on partition 1. It is also noteworthy that the models trained in age-order do not achieve the same level of syntactic categorization performance during training on partition 2 that was achieved by the models trained in reverse age-order which were trained on

partition 2 first. It appears as if the knowledge acquired by the SRNs during training on partition 1 is interfering with their ability to achieve the same level of syntactic categorization performance that was achieved during the first half of training by the SRNs trained in reverse-age order. Another remarkable finding is that syntactic categorization performance is slowly decreasing after an initial peak in performance achieved very early during training. This suggests that syntactic categories are acquired very early during training, and that subsequently acquired knowledge is slowly replacing knowledge about syntactic categories. It is possible that as finer-grained *semantic* distinctions are learned, the model does not need to hold on to more abstract knowledge about syntactic category membership, which is useful only for predicting a large category of words, rather than specific words. What about end-of-training performance? Does the early advantage for models trained in reverse age-order persist until the end of training? A t-test comparing end-of-training balanced accuracy across the two conditions reveals that it is not, $t = 1.14$ ($p = 0.16$). Thus, age-ordering appears to facilitate end-of-training semantic but not syntactic categorization performance.

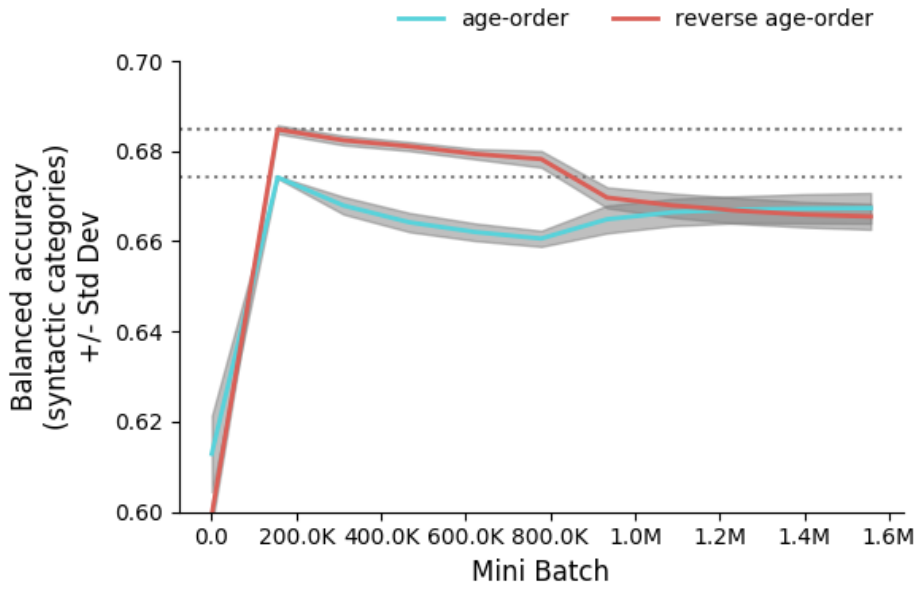

Figure 3.5
Average balanced accuracy as a function of training time (in minibatches) of 8 SRNs trained in age-order (blue line) and 9 SRNs trained in reverse age-order (red line) on 2 partitions of AO-CHILDES. The probe words used in this analysis are not the same 532 probe words used above, but are 830 word that can be unambiguously classified into one of eight syntactic categories (adjective, preposition, adverb, determiner, interjection, noun, pronoun, verb).

Conclusion

Age-ordered training of SRNs on AO-CHILDES facilitated acquisition of semantic, but not syntactic categories. The latter is in agreement with previous findings from Rohde & Plaut (1997). Rohde & Plaut convincingly argued that a reduction of the grammatical complexity of the input should not facilitate grammatical mastery because the SRN that 'starts small' has learned a skewed distribution of the target language that it is being tested on. A recent study makes a very similar point, by illustrating that 'the statistical evidence a corpus provides in favor of the target language falls off as its complexity deviates from the complexity of the language' (Rafferty & Griffiths, 2010). Simply put, when the task is to learn the grammatical rules underlying a sequence of symbols, whether in natural or artificial language, it is detrimental to

withhold the full range of grammatical complexity, even if only during the early stages of training. But why did semantic categorization performance benefit from training on age-ordered input? What exactly differentiates semantic from syntactic categories in the eyes of the SRN? I hope to provide answers to both questions in the subsequent chapters.

CHAPTER 4: AO-CHILDES STARTS SMALL

In the previous chapter I demonstrated that age-ordered presentation of sequences in AO-CHILDES facilitates semantic categorization in a simple recurrent network (SRN) compared to training in reverse age-order. I called this effect the age-order effect. Naturally, an explanation of the age-order effect will require an analysis of the incremental organization of the input. In this chapter I describe the results of several analyses comparing various corpus-statistical properties of partition 1 (speech to younger children) with partition 2 (speech to older children).

What are the building blocks of semantic categories?

When Elman found a beneficial effect of order on syntax acquisition, it was in the context of testing his 'starting small' theory. According to his theory, the SRN can better predict sequences consisting of complex embedded clauses when it is first trained on simple sentences with no embedded clauses before being trained on the full range of sentences. Syntax acquisition, according to Elman required learning a complex function that can be broken down into smaller components, which are themselves functions. Thus, learning the building blocks first, by training on input with fewer embedded clauses, would enable the SRN to learn a complex compositional function that it could otherwise not acquire[6]. For Elman, who was interested in syntax acquisition, one can easily point to the properties of the input that might enable the SRN to learn such building blocks (e.g. the proportion of sentences with embedded clauses). But, because I am concerned with semantic category learning, this question is not so straightforward. In fact, it is not even clear what the building blocks of semantic categories are.

[6] Since first proposing this idea, Elman has argued that the learner, not the input 'starts small'.

There are two directions to explore this question. First, it is possible that (distributional) semantic category learning - in the SRN or the infant - does not require knowledge about word-order or co-occurrence distance information. Instead, semantic categories may be understood as distributions over bag-of-words, where order and distance relationships are discarded. If true, it would not be necessary to conduct corpus analyses in which such information is preserved. Alternatively, word-order and co-occurrence distance might play an important role in (distributionally) defining semantic categories. If so, I would have to conduct corpus analyses that preserve this kind of information. Because it is not known what kind of information children actually use when constructing semantic categories[7], I decided to conduct both types of analyses.

Moreover, I will investigate not only the distributional properties of probe words, and how they might vary with age of the target child, but also the surface structural properties of the corpus as a whole. The probe word representations that the SRN acquires are not immune to learning experiences in which probe words do not occur. Therefore, it is possible that some incremental change in AO-CHILDES unrelated to probe words, like the overall syntactic complexity, could influence semantic category learning.

Speech to children starts small

What is already known about the incremental structure of the linguistic input to children? One way to answer this question is to compare child-directed to adult speech. A key finding in this literature is that speech to children is less lexically diverse compared to adult speech (Kirchhoff & Schimmel, 2005). Moreover, Foushee et al. (2016) found that lexical diversity gradually increases over the first three years of life before merging with adult-level lexical

[7] Provided that children use distributional cues at all to construct semantic categories.

diversity soon after. This means that a gradient in lexical diversity should also be present in AO-CHILDES. Similarly, it has been known for a long time that syntactic complexity, as measured via mean-length-of-utterance (MLU) is reduced in speech to younger children (Broen, 1972; Snow, 1972; Fernald & Morikawa, 1993). Another important difference between CDS and adult speech is that nouns are used more frequently and tend to refer to more concrete objects (Tardiff et al., 1997). Whether incremental changes in any of these factors can benefit semantic category learning from distributional information has not been previously investigated.

It is important to note that most of the beneficial properties of CDS on language acquisition are not captured by the text-based representation of AO-CHILDES, and are prosodic in nature (Fernald et al., 2006). Correspondingly, the most commonly reported benefits of CDS are on vowel discrimination (Trainor & Desjardins, 2002), word recognition (Singh, Nestor, Parikh & Yull, 2009), and speech segmentation (Nelson et al., 1989; Thiessen *et al*., 2005). Moreover, unique social factors related to CDS have been shown to influence language learning (Ramírez-Esparza, 2014). While I will not (and cannot) analyze these factors in AO-CHILDES, they are worth mentioning because they represent additional domains in which effects of the order of the input on learning can (and have been shown to) occur.

What incrementally changing corpus-statistical variables are most relevant for explaining the age-order effect? Naturally, I will look for any systematic differences between the distributional properties of probe words in partition 1 and partition 2. But what about syntactic complexity? Because syntactic complexity was shown to correlate with the age, I have chosen to also evaluate how syntactic complexity changes incrementally across AO-CHILDES. I say ‘estimate’ because I will use measures like MLU, and the number of unique sequences (of various sizes), which are at best indirect quantifiers of syntactic complexity. For simplicity, I will

use the term 'syntactic complexity' to refer to the complexity of surface-form structure as indicated by these measures. Concretely, I do not employ measures of syntactic complexity that involve linguistic knowledge, such as the number of discontinuities, embedded clauses, or inversions. Not only are indirect measures of syntactic complexity widely used, but more direct measures seem unwarranted given that the SRN has no a priori knowledge of English syntax.

After showing how syntactic complexity varies with the age of the target child in AO-CHILDES, I investigate whether the amount of information about semantic category structure changes incrementally. Specifically, I studied probe word contexts in partition 1 and 2 of AO-CHILDES to understand whether there are any systematic differences in the quantity of information about the semantic category structure. By 'semantic category structure' I mean the structure defined by the 532 hand-selected probe words and their membership in one of 28 semantic categories. Clearly, English contains semantic structure beyond these 532 probe words. But this 'true' semantic category structure underlying English is unknown and can therefore not be evaluated. These analyses would be most informative if I were to find differences in *either* syntactic complexity *or* the quantity of information about semantic category structure, but not *both*. If differences were observed along both dimensions, the results would not prove helpful in narrowing down and explanation of the age-order effect to a syntactic or semantic origin. On the other hand, if the syntactic changes across AO-CHILDES turn out to be much more prominent than any changes in the amount of information about the semantic category structure, I would conclude that *syntactic* complexity can influence acquisition of *semantic* categories in the SRN.

A question that is frequently asked in the context of child-directed speech and the early learning environment of children is whether speech to younger children is *supposed* to facilitate aspects of language learning. While it is plausible that caretakers (knowingly or not) modify their

speech according to the learners' abilities, I will not attempt to answer this question. Even in the face of numerous studies showing that caretakers tailor their speech to the abilities of the learner, this does not imply that such modifications are a required component of language acquisition. Moreover, the answer to this question may vary as a function of the language and therefore requires a cross-lingual approach.

A final note before I describe the corpus analyses: This chapter is not another investigation about how child-directed speech is unique. Instead, I ask: what are the properties of child-directed speech, that can be captured by a text-based representation, that *change as a function of age of the target child*? The question of how CDS is a simplification of adult-speech is a clearly related question, but it is not the *same* question. The distinction is important because not all child-directed speech in AO-CHILDES is necessarily a simplification of adult speech. I use 'child-directed' to refer to any speech that is directed at a child, rather than speech that is simplified *because* it is child-directed.

Syntactic Complexity

Syntactic complexity can be estimated by a number of factors, which I will explain shortly. In subsequent chapters I will often refer to 'syntactic complexity' simply as 'complexity' in order to remind the reader that these measures quantify the complexity of the surface-form structure of the input, and not aspects of any deep-structure (e.g. syntactic parse trees). There at least two reasons why syntactic complexity is relevant to learning semantic categories. First, even if the difficulty of the *task* (e.g. semantic categorization) remains constant, a more complex *learning environment* (e.g. syntactic complexity) may negatively impact learning of the task. For example, a more complex learning environment would place greater strain on a finite-capacity

representational system. Representation of more task-irrelevant information (e.g. syntactic relationships between words) can easily crowd out knowledge about task-relevant information (e.g. semantic categorization). Second, syntactic complexity may introduce additional variation into the distributions defining semantic category membership of probe words. By so doing, increasing the syntactic complexity would reduce the quantity of information about semantic category structure in the input. For example, the introduction of novel constructions may decrease the likelihood that two semantically related probe words occur in similar contexts. Specifically, longer utterances and more varied use of vocabulary words and multi-word constructions could drive apart the similarity of probe word contexts. Decreasing the similarity between contexts of semantically related probe words would impair distributional learning. If true, one might interpret the negative effect of increased complexity on semantic category learning as being *mediated* by a reduction in the amount of information about semantic categories. I will investigate the following indices of syntactic complexity: the distribution of syntactic categories, Shannon entropy, mean utterance length, word repetition, the Taylor exponent, n-gram model perplexity, and the number of unique constructions (n-grams of orders 1-7).

Distribution of syntactic categories

Syntactic categories for each vocabulary word were obtained by POS-tagging AO-CHILDES with Python module *spacy*. The syntactic category assigned to a vocabulary word is the category which was most frequently assigned to all occurrences of the word. To understand how the distribution of syntactic categories changes across the input, AO-CHIDLES was split into 256 equally sized partition, and the proportion of nouns, verbs, adjectives, prepositions,

determiners, pronouns, and punctuation was quantified. For each category, a vector of 256 proportions was obtained and correlated (using Spearman's correlation) to the rank of the partition (indicating its position in AO-CHILDES). A positive correlation indicates that a syntactic category is more representative of the speech to older children, while a negative correlation indicates that a syntactic category is more representative of speech to younger children.

I found that the number of prepositions (rho=0.66) and pronouns (rho=0.71) correlate most positively with age. Of those that correlate negatively with age, the number of punctuation (rho=-0.58), nouns (rho=-0.67) and determiners (rho=-0.44) are largest. Because the number of determiners and nouns in a partition are positively correlated (rho=0.59), it is likely that the reduction in both nouns and determiners with age is due to the strong association between the two syntactic categories. What can we take away from these results? One, I have confirmed that speech to younger children is more noun-rich than speech to older children. Two, the negative correlation between punctuation and partition rank (a proxy for age) confirms that utterances spoken to younger children are shorter than utterances directed at older children. This also confirms the well known fact that caretakers pause more frequently when speaking to younger children.

The correlations between *predictor* variables, revealed a very clear trade-off between punctuation and prepositions (rho=-0.79), and punctuation and conjunctions (rho=-0.64). The negative correlations indicate that punctuations is gradually replaced by prepositions and conjunctions as a function of age. This must be true, given that in English the addition of conjunctions or prepositions typically extends the length of a sentence. Interestingly, the increase in pronoun density is negatively correlated with the increase in noun density (rho=-0.62) which

suggests that nouns - mostly proper nouns - are gradually replaced by pronouns in speech to older children.

Shannon entropy and mean utterance length

I computed the Shannon entropy (also a measure of lexical diversity) of the discrete distribution of word frequencies in each of the 256 partitions, and plotted the results as a function of partition rank (indicating the location in AO-CHILDES, and a proxy for age) in figure 4.1. I also included a plot of the mean utterance length, and the standard deviation of the utterance length. While the figure shows there is a considerable amount of variance, a clear upward trend can be discerned for each factor. The most pronounced upward movement is for the standard deviation of utterance length, which appears to be steadily increasing until about the midpoint in AO-CHILDES. The upward movement of the Shannon entropy, on the other hand, is primarily restricted to the first quarter, or less, of AO-CHILDES. Underlying the curve for the mean utterance length, a steady but gradual upward movement is detectable. The mean length of utterance for the first half of AO-CHILDES is 5, while the mean utterance length of the second half of AO-CHILDES is just short of 8.

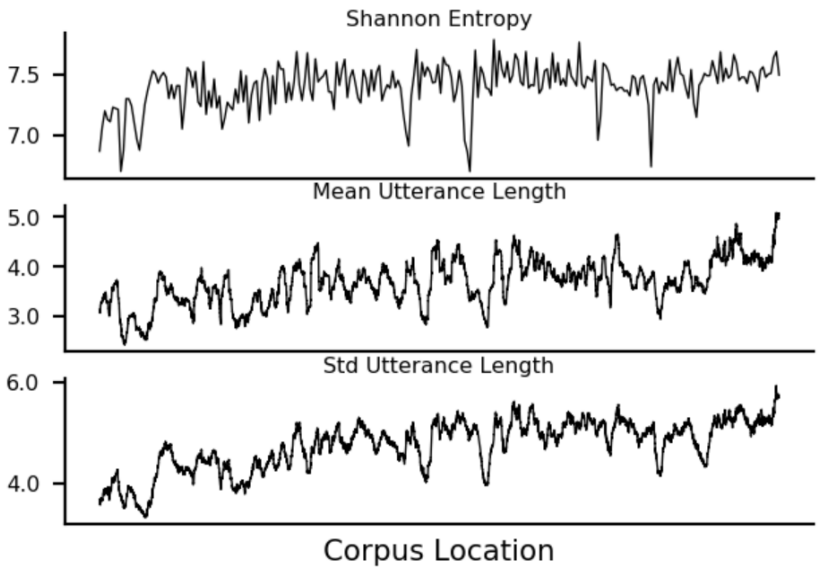

Figure 4.1
Three measures of syntactic complexity plotted as a function of the location in AO-CHILDES (rank of partition, 1-256).

Word Repetition

Frequent reuse of the same word makes for a less complex learning environment. In that light, I obtained the 100 earliest and 100 latest occurring vocabulary words, and plotted their frequency as a function of location in AO-CHILDES (partition 1-256). Are the earliest used words repeated more often than the words used latest in AO-CHILDES? The results are shown in figure 4.2. The answer is a clear yes; the maximum total frequency of the 100 earliest words (shown in orange) is larger than 3,000, whereas the maximum total frequency of the 100 latest occurring words is just short of 1,000. I have confirmed similar trends with different set sizes.

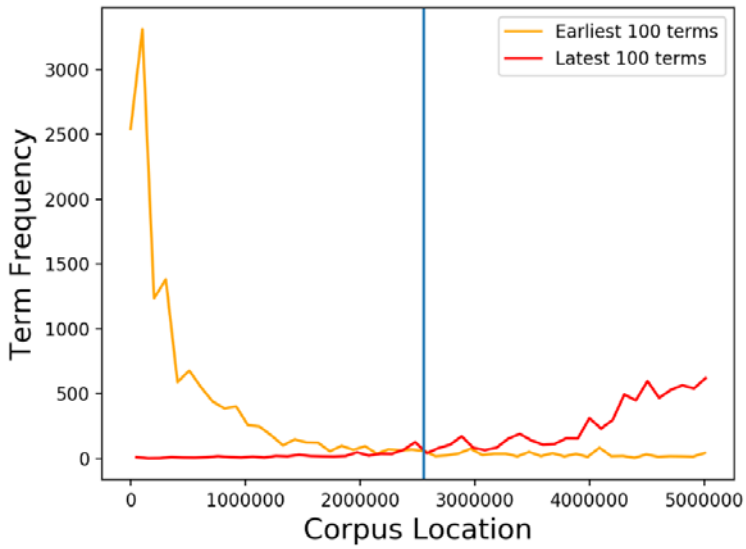

Figure 4.2
Total frequency of words occurring earliest (orange line) and latest (red line) in AO-CHILDES as a function of corpus location (partition 1-256). The midpoint location is indicated by the vertical blue line.

The Taylor exponent

A recently introduced measure of structural complexity of linguistic sequences is the Taylor exponent (Kobayashi & Tanaka-Ishii, 2018). It is the exponent in a power-law relationship between the variance of word frequency and the average word frequency per unit of time. This relationship, known as Taylor's law, was first discovered in ecology where the variance of the number of individuals of a species per unit area is related to the average number of individuals per unit area according to a power law. Taylor analysis has since been applied in numerous other fields (Eisler, Bartos, and Kertész, 2007) to demonstrate systematic relationships between events. Theoretically, an independent and identically distributed (iid) process must have a Taylor exponent of 0.5, and larger exponents indicate processes in which events depend on each other. Human linguistic sequences exhibit a Taylor exponent above 0.5 (Kobayashi &

Tanaka-Ishii, 2018), indicating that words co-occur systematically. Moreover, CDS is characterized by a larger Taylor exponent than adult speech, which suggests that CDS is more template-like. Because I am interested in category learning from distributional information, I wanted to know whether speech to younger children is more systematic, and less structurally complex than speech to older children.

I computed the Taylor exponent separately for partition 1 and 2 of AO-CHILDES using the same method used in (Tanaka-Ishii & Kobayashi, 2018). FirstI split each partition into chunks of 5,600 words and computed the frequency of all words in each chunk. Then I obtained the average and standard deviation of the frequency of each word across the chunks and fitted the resulting data to a linear function in log-log coordinates by the least-squares method. The results are shown in figure 4.3. The relationship between the standard deviation and mean of each word's frequency is shown for words in partition 1 (lower left panel) and partition 2 (lower right panel). The Taylor exponent, indicated on each plot, represents the slope of the best fit line in log-log coordinates. The Taylor exponent associated with partition 1 (0.631) is larger than the Taylor exponent associated with partition 2 (0.612). This indicates that partition 1 contains a greater number of constructions with fixed forms compared to partition 2. As mentioned before, a Taylor exponent calculated for an 'iid' process is 0.5, and the larger the value the more template-like it is. This finding is consistent with that of Kobayashi & Tanaka-Ishii (2018) who found that CDS speech (in addition to programming languages and music) was found to have a

larger Taylor exponent than adult language.

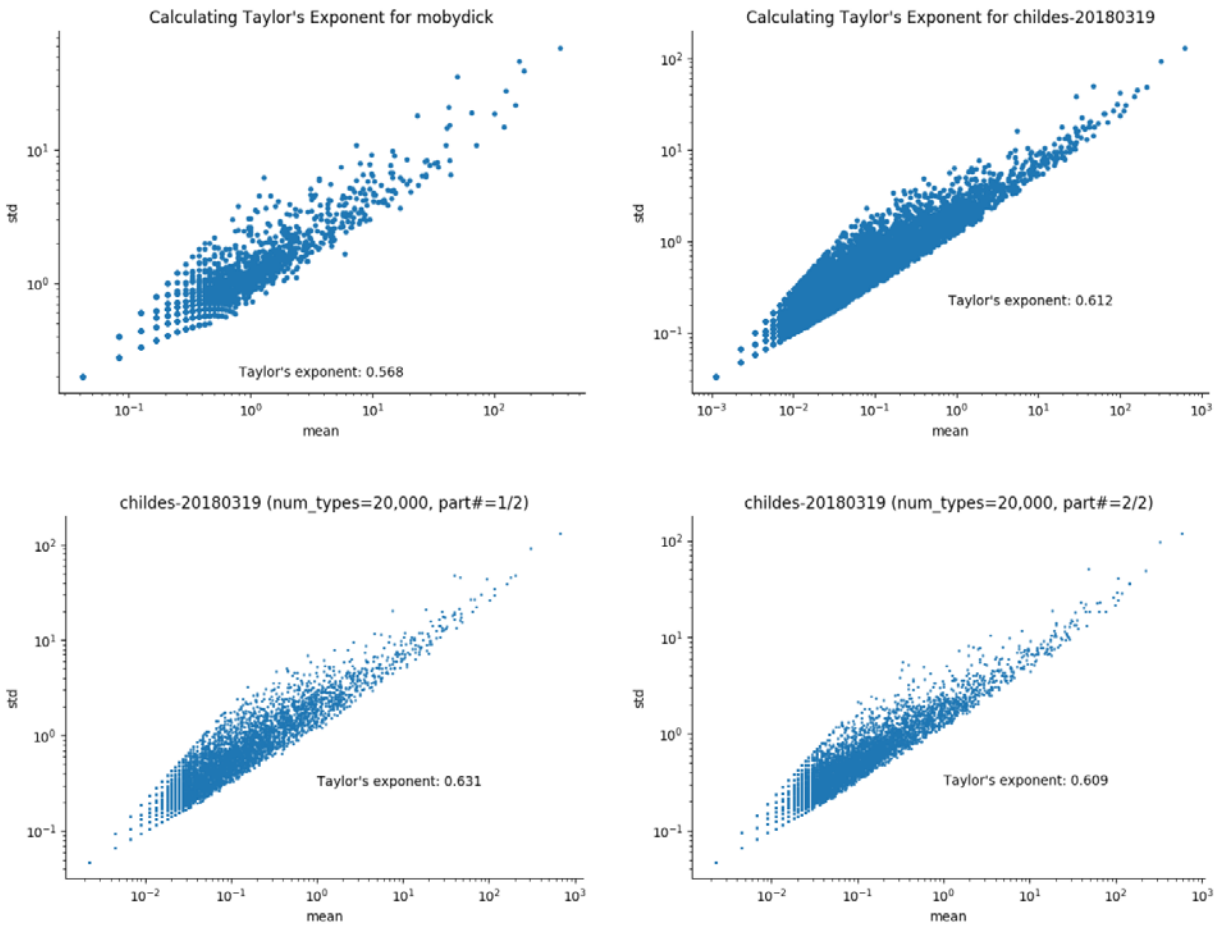

Figure 4.3
Scatterplots used to compute the Taylor's exponent for the novel 'Moby Dick' (top left panel), AO-CHILDES (top right panel), partition 1of AO-CHILDES (lower left panel), and partition 2 of AO-CHILDES (lower right panel). Each point represents the relationship between the standard deviation (y-axis) and the mean (x-axis) of a word's frequency. Plotted in log-log coordinates.

N-gram model perplexity

No analysis of linguistic complexity would be complete without evaluating n-gram language model fits. Specifically, I split AO-CHILDES into 2 equal sized partitions and trained Kneyser-Ney language models of varying n-gram sizes (3 to 6) separately on each partition. I performed this analysis twice. In one analysis, I trained n-gram models on input where any out-

of-vocabulary word had been replaced by an out-of-vocabulary symbol. I repeated the analysis with the full vocabulary (all words left intact). N-gram language models were trained with the KenLM Language Model Toolkit (Heafield et al., 2013) and scored using the python module *kenlm*. The results are shown in figure 4.4. Using the reduced vocabulary (all but the 4,096 most frequent words), average perplexity scores for all three n-gram sizes were smaller when trained and evaluated on partition 1 compared to partition 2 (4-grams: 9.1 vs 9.6, 5-grams: 5.4 vs 5.9, 6-grams: 4.2 vs 4.6). The same pattern is observed when training n-gram language models on partitions with the full vocabulary (4-grams: 8.8 vs 9.4, 5-grams: 5.4 vs 5.8, 6-grams: 4.2 vs 4.6). As mentioned before, perplexity is a measure of sequence prediction error. In that sense, a lower perplexity indicates that it is easier to predict the next word given information about the words that come before it. Perplexity can also be viewed as a measure of how unlikely a model judges a sequence of words, given its overall training experience. In all six cases, an n-gram model judges sequences in partition 2 to be less likely than an equivalent n-gram model trained on partition 1. One interpretation is that there are more word sequences in partition 2 than in partition 1 that conform to the model's expectations. Again, this confirms that partition 2 is more complex than partition 1.

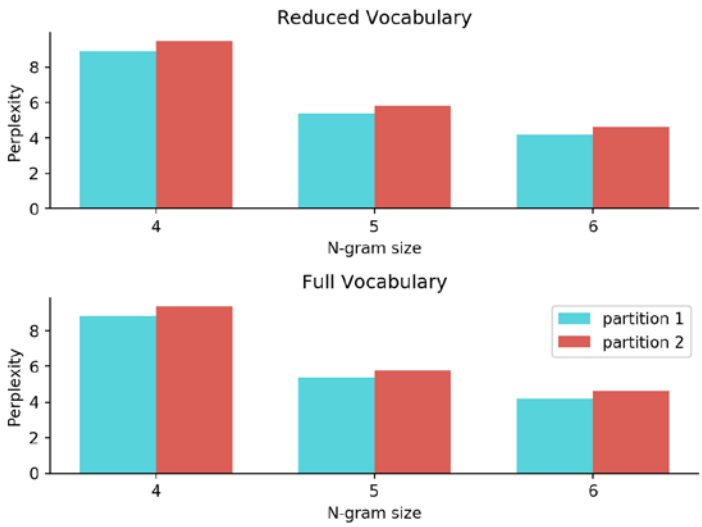

Figure 4.4
Perplexity of various n-gram models trained on partition 1 (blue) and partition 2 (red) of AO-CHILDES. N-gram models trained on input where all but the most frequent 4,096 words were replaced by an out-of-vocabulary symbol (top panel) and on input with all words left intact (lower panel)

Number of unique constructions

A final measure of syntactic complexity is the number of unique constructions. A greater number of unique constructions can indicate a number of phenomena, including drawing from a larger vocabulary of words (e.g. talking about a broader set of topics), using longer utterances (e.g. multi-clause utterances, multiple noun phrases, more frequent use of prepositional phrases, insertion of adjectives and/or adverbs), and greater usage of less frequent or alternate constructions (e.g. past tense). It could also indicate that the content of speech is specified more clearly in terms of meaning (e.g. tense and aspect). Not all of the above are syntactic in nature, but most of them are. But what is a construction? I define a construction here as a sequence of words (also known as a bi-gram). In my analysis, I counted the number of unique n-grams of size 1 through 7 to maximize the likelihood of detecting any differences between speech to younger and speech to older children. I conducted two analyses: In the first, I obtained all unique n-grams

of a specific size that are in AO-CHILDES. Next, I asked what percentage of those unique n-grams are found in partition 1 and what percentage of those unique n-grams are found in partition 2. Put differently, I calculated the percentage of unique n-grams in both partitions that are in partition 1, and the percentage of unique n-grams in both partitions that are in partition 2. The results of this analysis are shown in the left panel of figure 4.5. I found that partition 2 captures a greater proportion of the total number of unique n-grams (of size 2 through 7) compared to partition 1. For example, nearly 70% of the total number of unique 2-grams can be found in partition 2, while only 65% can be found in partition 1. In a second, related analysis, I asked what percentage of n-grams in a partition are also found in the other partition. This measure isn't strictly speaking a measure of syntactic complexity, but an indicator of the potential that training on one partition may generalize better to the other. If it is true that partition 1 facilitates subsequent learning, one way this could be explained is that constructions seen during training on partition 1 recur more frequently during training on partition 2 than the other way around. If such results were to be found, this would indicate that a greater proportion of the constructions in partition 1 are more 'foundational' or 'typical' and are therefore reused more often. The results are shown in the right panel of figure 4.5. Indeed, I found a small advantage for partition 1. The proportion of 2-grams in partition 1 that recur in partition 2 is slightly larger compared to the proportion 2-grams in partition 2 that recur in partition 1. But this difference does not hold for n-grams of larger sizes.

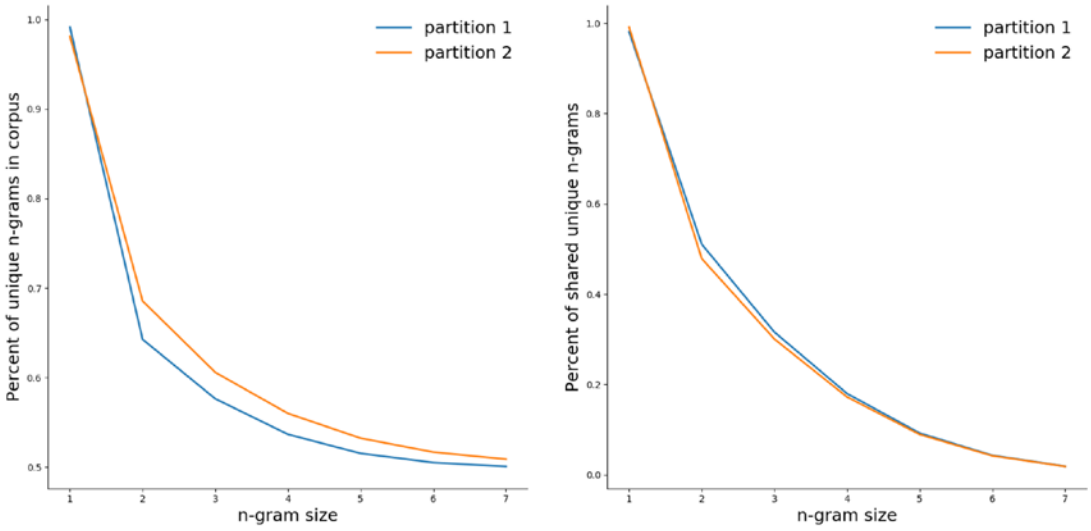

Figure 4.5
Quantifying n-grams in AO-CHILDES. Left panel: Percent of unique n-gram sin partition 1 (blue) and partition 2 (orange). Right panel: Percent of unique n-grams in partition 1 (blue) that also occur in partition 2, and percent of unique n-grams in partition 2 (orange line) that also occur in partition 1. N-gram sizes used are 1 through 7.

Information about Semantic Categories

All of the analyses described above point at the presence of a syntactic complexity gradient in AO-CHILDES. Assuming that the age-order effect is in part caused by the reduced complexity in partition 1, I asked whether the effect of the reduced complexity of partition 1 is mediated by an increase in the availability of distributional information about the semantic category structure defined by the 532 probe words. I did not undertake a mediation analysis, but reasoned that if I were to find a difference in information about semantic categories between the two partitions, it would *likely* be driven by the gradient in syntactic complexity. I don't think that the amount of distributional information about semantic categories in speech is a variable that is under independent control by the speaker. If so, this would mean that speakers compute distributional information when preparing utterances. This would be a costly procedure, and would most likely only benefit the listener - and only in restricted circumstances (e.g. when

speaking to children). To modify semantic category information, it could be accomplished more straightforwardly by modifying a more globally operating variable, such as syntactic complexity. To be more concrete, I provide an example of how increasing the complexity of an utterance can reduce the amount of distributional information about a probe word's semantic category membership. Assume a learner has already encountered all the utterances in AO-CHILDES, and hears either utterance (a) or (b). Because (b) has an additional, optional adjective, it represents an experience with high syntactic complexity, and (b) in turn represents an experience with relatively low syntactic complexity. In a high complexity language environment, utterance (b) is more likely to occur, because syntactic complexity is partly characterized by more frequent usage of adjectives. Previously, the context *Do you want some* has proven predictive of probe words in the category DRINKS. As such, utterance (a) represents a continuation of this trend, and reinforces the informative link between the context and the category. On the other hand, the syntactically more complex utterance (b), contains an additional word between the informative context and the probe word. Assuming that *Do you want some* has been perfectly predictive of probe words in the category DRINKS, the fact that it is followed by *more* in (b) has reduced its informativeness. *Do you want some* is no longer perfectly predictive of DRINKS, because it can also be followed by *more*.

(a) Do you want some {juice, milk, coffee, ..}?
(b) Do you want some more {juice, milk, coffee, ..}?
(c) Do you want some additional {juice, milk, coffee, ..}?

Utterance (c) represents a different learning experience in a complex language environment. Syntactic complexity is higher relative to utterance (a), but is identical to utterance (b). The only difference is in the choice of adjective. In a complex language environment such as

partition 2 of AO-CHILDES, not only does syntactic complexity increase, but lexical diversity increases too. This means that components of semantically informative contexts are more likely to be replaced by alternative words that serve a similar function. Thus, increased lexical diversity, too, can reduce the strength of a context-category link. If utterance (b) and (c) *both* occur, then the strength of the link between the context in (b) with DRINKS and the context in (c) with DRINKS is weakened compared to if *only one or the other* context occurred. The two contexts retain their informativeness, but the SRN has less experience with each.

What exactly happens in an SRN, trained on child-directed speech that includes utterance (a), when it first experiences utterance (b) or (c)? Does the SRN treat the additional adjective *more* as noise? Does the SRN learn to ignore the adjective, and re-represent the context *Do you want some* such that it is still predictive of DRINKS but at a greater distance? Perhaps it learns that the context can be predictive at multiple distances? Alternatively, the SRN might learn a new context, *Do you want some more*. This context may be represented separately of *Do you want some*, and there would be no need to modify the existing representation of *Do you want some*. If so, one might ask whether semantic categorization performance, in general, suffers when there are a larger number of informative contexts compared to a few contexts that are equally informative. Given finite representational capacity, it is plausible that the SRN would benefit from experiences with a small number of contexts each of which occurs more frequently compared to a large number of equally informative contexts each of which occurs less frequently. A related question is whether increasing the number of context-category links generally results in more or less informative context-category links. On the one hand, a larger number of context-category links reduces the frequency with which each occurs; on the other hand, the chance that two unrelated probe words occur in identical or similar contexts is reduced.

The latter possibility indicates that more complex input could actually increase the amount of distributional information about semantic category structure. Answering all these questions is beyond the scope of this work.

The analysis conducted here is focused on the following three questions: How substitutable are same-category probe words in partition 1 compared to partition 2? How much information about semantic category structure is captured by a bag-of-words model trained on partition 1 compared to partition 2? Similarly, how much information about semantic category structure is captured by a term-by-window (sliding window) model trained on partition 1 compared to partition 2?

Same-category Probe Substitutability

The goal of this analysis is to find any differences in how substitutable same-category probe words are in partition 1 compared to partition 2. While distributional learning is typically defined as learning the similarity of the contexts in which entities occur, it can be recast as learning how substitutable entities are in the contexts in which related entities occur. Successful category learning given only distributional information requires that category members must occur in the same or similar contexts; otherwise there would be no reason to group category members. For example, if the words *cat* and *dog* are highly substitutable given the contexts in which they occur, then a distributional learning system would assign similar representations, and would therefore be more likely to group them into the same category. If same-category probe word substitutability were greater in partition 1 compared to another, this would indicate that the information about semantic category structure is greater in partition 1 compared to partition 2.

First, I obtained all probe word contexts in partition 1 and all probe word contexts in partition 2. Contexts are defined here as a sequence of words that are left-adjacent to a probe word. I included contexts with size 1 through 4. For each context, I constructed a probability distribution over probe words. This distribution represents how likely a probe word occurs given that context. Because a context can be associated with any probe word, a single context is often associated with multiple categories. For each category associated with a context, I computed the Kullback-Leibler (KL) divergence between the observed probability distribution (over probe words) with the expected probability distribution (over probe words) had the context been exclusively associated with a single category, and equally with all members of the category. The KL divergence is a measure of how much additional information is captured by the observed distribution that is not captured by the expected distribution. Simply put, the KL divergence is a measure of the dissimilarity of two probability distributions. The higher the KL divergence, the less similar are the observed and the expected distributions, and this means that the context is not a good cue to membership in the category under question. Because the expected probability distribution assumes that all probe words in a category should equally likely occur in the context, a higher KL divergence indicates that probe words in the same category are not substitutable for one another. I calculated the KL divergence for all context-category pairs for partition 1 and partition 2 separately, and plotted the results in figure 4.6. Each panel shows a frequency-normalized histogram of all KL divergences obtained for partition 1 (in blue) and partition 2 (in red). If same-category probe words are more substitutable in partition 1, the histogram shown in blue should be more left-skewed (indicating overall lower KL divergences) than the histogram shown in red. Across all context sizes 1 through 4, I found the KL divergences to be slightly smaller. The left-skew of the histogram representing partition 1 (in blue) is not obvious, so I have

shaded the areas blue where it is assigned higher probability values than the histogram representing partition 2 (in red). Notice that the shaded regions are consistently associated with below-average KL divergence. This means that a slightly larger number of contexts in partition 1 in which same-category probe words are more substitutable. Given the large number of context-category pairs (KL divergences) a two-sample independent t-test comparing the mean of the KL divergences obtained from partition 1 to the mean of the KL divergences obtained from partition 2 is highly significant ($p < 0.0001$ for all context sizes). These findings suggest that there are a *very small* number of cases where partition 1 provides more distributional information about semantic category structure (e.g. same-category probe words are more substitutable).

Lastly, a note of caution: it is not known whether probe substitutability is measuring something that the SRN is actually sensitive to during training. In fact, this is true of any of the corpus analyses conducted here. Whether the small difference in same-category probe word substitutability detected here is therefore relevant for explaining the age-order effect requires more direct testing.

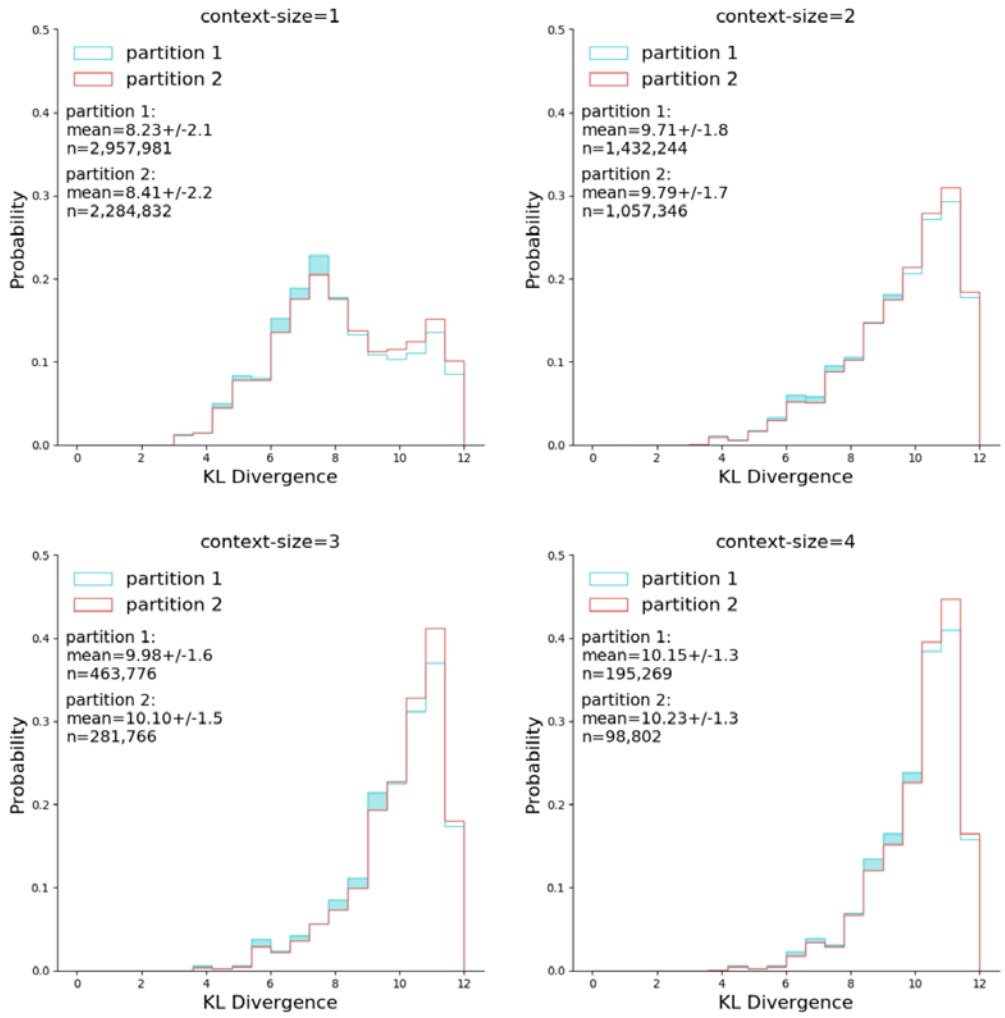

Figure 4.6
Same-category probe word substitutability (measured in KL divergence) for partition 1 (red) and partition 2 (blue). Context size increases clockwise from top left.

Balanced Accuracy of AO-CHILDES bag-of-words model

A more direct way to compute how much distributional information about semantic category structure is available is to use the same measure used to compute semantic categorization performance in the SRN, the balanced accuracy. Instead of computing the balanced accuracy for probe word representations learned by the SRN, I computed the balanced accuracy for probe word representations ‘learned’ by a bag-of-words model that received as

input either partition 1 or partition 2. The bag-of-words model represents a word as a vector where each element is the frequency with which a vocabulary word occurs with the word in question. The co-occurrence context is restricted to a sequence of words occurring left to the word in question. The sequential information, however, is discarded by updating the frequency counter of a context word regardless of its position in the sequence. Similar to how I evaluated the SRN, I tracked the balanced accuracy at multiple timepoints, starting when the bag-of-words model has not received any input, and ending when it has seen all of the input in a partition. If the balanced accuracy of the bag-of-words model 'trained' on partition 1 is greater (across all timepoints) this would support the idea that partition 1 has more distributional information about the semantic category structure than partition 2.The results, shown in figure 4.7, demonstrate that this is not the case. Across all context sizes tested (1-4 shown, and also 4-7), the balanced accuracy consistently rises faster and reaches a higher endpoint for the bag-of-words model 'trained' on partition 2. This finding would lead to the opposite conclusion drawn from the analysis described above, in which probe substitutability was found to be slightly larger for category contexts in partition 1. Another interesting trend shown in figure 4.7 is that the gap in balanced accuracy between the bag-of-words model trained on partition 1 and 'trained' on partition 2 is increasing proportionally to the context size. Specifically, the balanced accuracy obtained by the bag-of-words model 'trained' on partition 2 drops less quickly with increasing contexts size than the model 'trained' on partition 1. One way to interpret this finding is that the information about semantic category membership is preserved across greater distances in partition 2 compared to partition 1. This is likely related to the fact that utterances tend to be longer in partition 2, and that semantic dependencies can occur therefore span longer distances in partition 2.

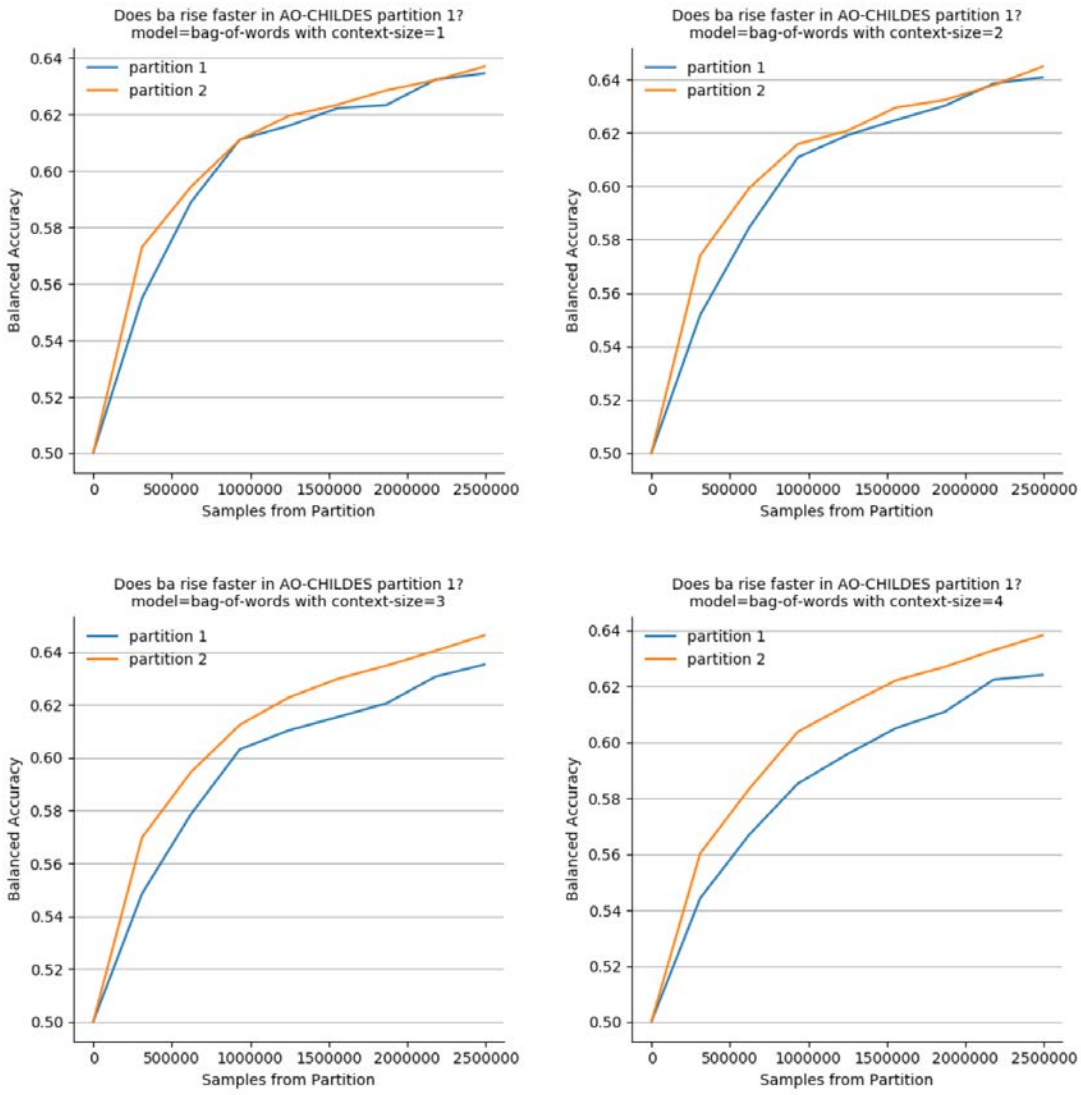

Figure 4.7
Balanced Accuracy computed on probe word representations acquired by a bag-of-words model with partition 1 as input (blue) and partition 2 as input (orange). Probe word representations are vectors where each element represents the frequency with which a vocabulary word occurs in the left context of a probe. Various sized contexts are used; context size increase in clockwise order starting with the top left panel.

Balanced Accuracy of AO-CHILDES term-by-window model

An alternate representation of probe words in which information about word order is preserved can be captured by a term-by-window model. In contrast to the bag-of-words model,

the term-by-window model slides a windows (1 through 4 used here) across the words in the input, and updates a vector of co-occurrence frequencies where each element represents both the identity of the word and it position in the sliding window. When the context size is 1, the bag-of-words model is a special case of the term-by-window model. When the context size is larger than 1, the size of each vector increases by a factor of the context size. Given a vocabulary of 4,096 words, a term-by-window model representation ‘trained’ with a context size of 1 is 4,096 elements long; when the context size is 2, the representation is 8,192 elements long. As described above, I ‘trained’ one model on partition 1 and another on partition 2, and tracked the balanced accuracy as a function of the amount of input each has seen. The results are shown in figure 4.8. The balanced accuracy tends to be larger for the term-by-window model ‘trained’ on partition 1 when the context size is 3 and 4, but this trend does not hold for context sizes 1 and 2. This indicates that there is some, but not much, more information about the semantic category structure in partition 1.

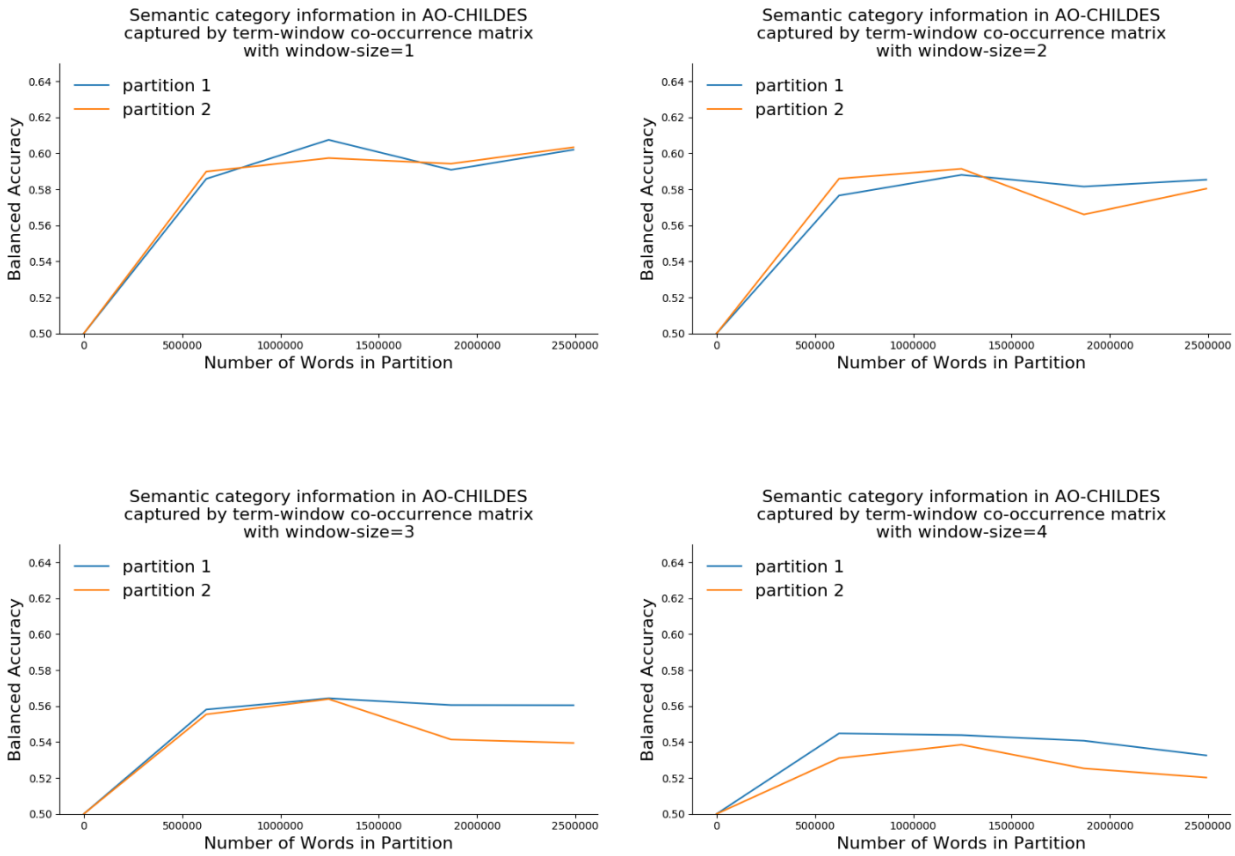

Figure 4.8
Balanced Accuracy computed on probe word representations acquired by term-by-window model with partition 1 as input (blue) and partition 2 as input (orange). Probe word representations are vectors where each element represents the frequency with which a vocabulary word co-occurs to the left of the probe word *at a specific distance*. Various sized contexts are used; context size increase in clockwise order starting with the top left panel.

An interesting trend in this set of results is that the balanced accuracy drops considerably faster for the term-by-window models compared to the bag-of-words models. In other words, adding word-order information to the probe word representations weakens semantic categorization performance. This means that the SRN's encoding of the sequential structure in AO-CHILDES might negatively influence semantic category learning. The best performing SRN should be the one that can learn to *ignore* the sequential structure, and focus only on semantic dependencies regardless of the distances they span. Clearly, this cannot happen in an SRN

explicitly trained to predict sequences. This kind of observation, however, suggests that the best performing model of distributional semantics cannot be the SRN, because it is too constrained by the sequential structure of the input. An alternative model that is not constrained in this way, Word2Vec, trained on the same input as the SRN, does perform better on semantic categorization (Huebner & Willits, 2018), but not by much. That said, it is remarkable that the SRN can capture semantic structure *despite* being strongly influenced by word-order, and *despite* not being explicitly trained to extract semantic relationships, as Word2Vec is.

Conclusion

The results of the probe substitutability analysis and the comparisons of the term-by-window models are in agreement with each other: In both cases, I concluded there is slightly more amount of information about the semantic category structure in partition 1 compared to partition 2. This makes sense because in both analyses word-order information was preserved. After word-order information was removed (bag-of-words model results), I found a small advantage for partition 2. What is the overall verdict? Is there more distributional information about the semantic category structure in partition 1? The answer seems to depend on the kind of distributional information in question. When distributional information is restricted to information that preservers word-order, then the answer appears to be yes. If this restriction does not apply, the answer appears to be no.

Is the difference in the amount of distributional information about the semantic category structure between partitions caused by the change in syntactic complexity between the two partitions? The corpus analyses cannot provide an answer to this question. However, the small effect size of the change in the amount of information about semantic category structure relative

to the effect size observed for changes in syntactic complexity suggest that the age-order effect is better explained by a theory in which syntactic complexity can *directly* influence semantic category learning, rather than *indirectly* by modifying the information available about the semantic category structure. Given the consistent results demonstrating a relatively strong difference in the syntactic complexity between partitions, it would be surprising if the gradient in syntactic complexity did not contribute to the age-order effect. While it is possible that the difference in syntactic complexity plays no role in the age-order effect, it would be even more surprising if the smaller and less consistent difference in the amount of information about semantic category structure did play a role.

Addendum: Locations of Words

An analysis unrelated to either syntactic complexity or semantic category structure concerns the location of words in AO-CHILDES. Due to the sequential nature of the training used to observe the age-order effect, it would be useful to know precisely which words tend to occur more frequently in partition 1 compared to partition 2. I have already shown that the syntactic categories are not evenly distributed across partitions; for example partition 1 is more noun-rich than partition 2. But what about the *content* of speech? Does the content change in any systematic fashion from partition 1 to partition 2? To answer this question, I obtained the cumulative frequency trajectory of each vocabulary word incremented at every one of 256 partitions of AO-CHILDES, and sorted them from most-decreasing to most-increasing[8]. In figure 4.9, I plotted the cumulative frequency trajectories of the 6 most-

[8] The actual criterion used was the slope of a best-fit line that best approximated the cumulative frequency trajectory.

decreasing (top panel) and 6 most-decreasing words. An interesting trend emerged from this analysis. The words that are used the most to speech to younger children are referential: *the*, *there*, and *here*. This in agreement with previous findings which found that speech to children tends to more restricted to the referential context. Utterance boundary markers too, were included among the words with the most decreasing frequency trajectory. Periods and exclamation marks, while silent in speech, are viewed by many researchers as playing an important role in advancing an infant's ability to segment the speech stream. The role that utterance boundary markers play in distributional learning after segmentation has been completed (the input to the SRN used here is segmented) is less clear.

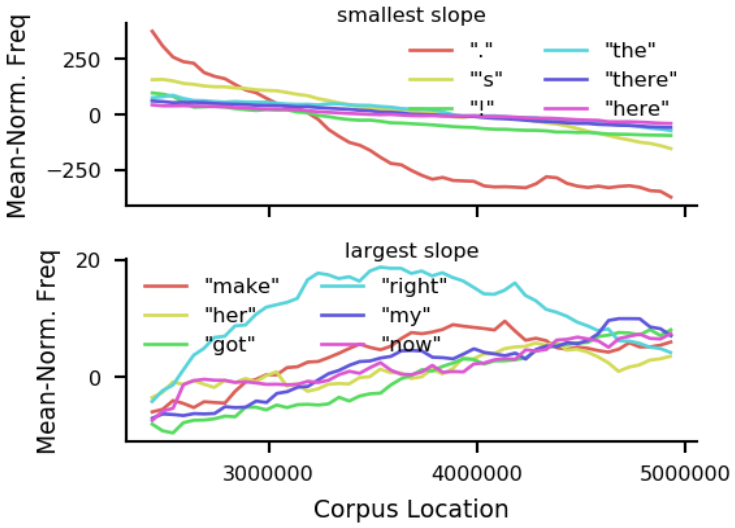

Figure 4.9
Cumulative frequency trajectories of various words across 256 partitions of AO-CHILDES. The top panel shows the trajectories for words which were classified as most-decreasing, and the lower panel shows trajectories for words which were classified as most-increasing.

In the analyses above I asked whether there is more distributional information about the semantic category structure (defined by the 532 probe words) in partition 1. The answer is not

straightforward, and more analyses need to be conducted. A different - though, more crude - way of answering this question is to simply count the number of probe words in each partition. Are there more probe words in partition 1? If so, the SRN would have a greater number of experiences to refine its representations for probe words. In figure 4.10, I plotted the total type frequency (top panel) and token frequency (bottom panel) of all probe words in partition 1, and partition 2. I further broke down the analysis by whether a probe word tends to occur more frequently in the first half (blue line, 'early probes') of AO-CHILDES or the second half (orange line, 'late probes'). An equal amount of probe word types occur in both sets (marked 'early probes' and 'late probes' in the figure). The dotted line represents the average between the two lines. It also indicates approximately how many 'early' or 'late' probe words are expected to occur in a partition had the two partitions come from identical distributions. The top panel reveals that nearly all probe word types that occur in partition 1 also occur in partition 2 (save for 1 probe word). The bottom panel is more revealing: 'early probes' occur more frequently in partition 1 by a respectable margin. Concretely, there are 13,132 more occurrences of 'early probes' than would be expected by chance (observed = 78,750 vs. expected = 65,618). Similarly, there are 7,893 more occurrences of 'late probes' partition 2 than expected had the two partitions been sampled from the same distribution (observed = 60,728 vs. expected = 52,835). Importantly, this suggests that information about semantic categories in one partition may not be the same information that is available in the other partition. The SRN, trained sequentially on both partitions, must therefore preserve the knowledge about semantic category structure gained during training on the first partition when training on the second partition. Otherwise, it risks forgetting information that is not available in the second partition.

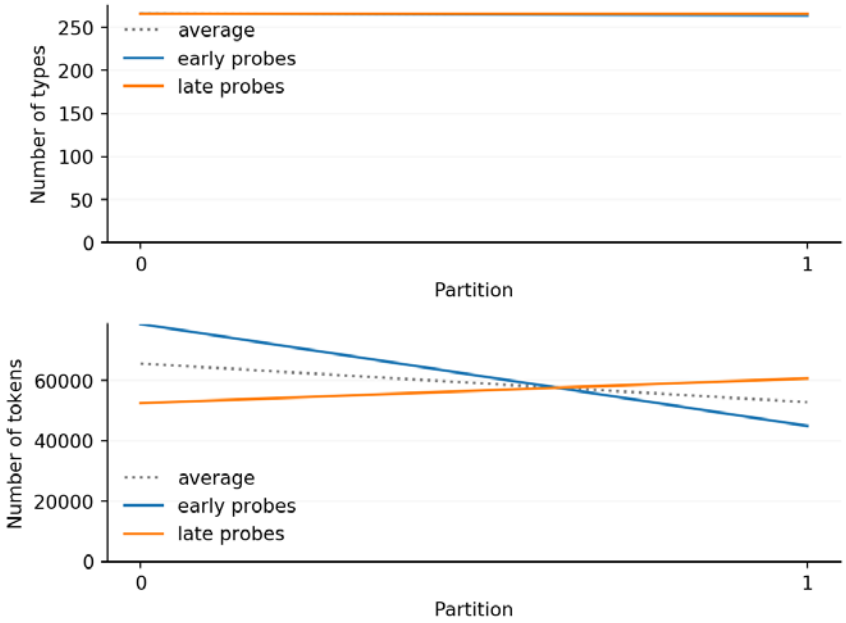

Figure 4.10
Type frequency (top panel) and token frequency (lower panel) of probe words that tend to occur more frequently in partition 1 (blue line) and probe words that tend to occur more frequently in partition 2 (orange line). An equal number of probe words is in both sets.

In light of these findings, it is possible that the content of speech changes with increasing age of the target child. For example, speech to younger children should involve topics more closely restricted to the referential context, and to basic interactions such as eating, and dressing. To verify that this is indeed the case, I plotted the cumulative frequency trajectories for the words *bottle* and *story*, which are shown in figure 4.11. Indeed, the word *bottle* is more likely to occur in speech to younger children, and the word *story* is more likely to occur in speech to older children. Younger children are more likely to require assistance from their mothers while being fed than older children, and the word bottle is more likely to occur in such circumstances. Older children, on the other hand, are more likely to be read to, and reading time is a circumstance under which the word *story* is very likely to occur. The cumulative frequency trajectories of other probe words

with either a strong increasing or decreasing trend do not lend themselves to interpretation as readily as do *bottle* and *story*.

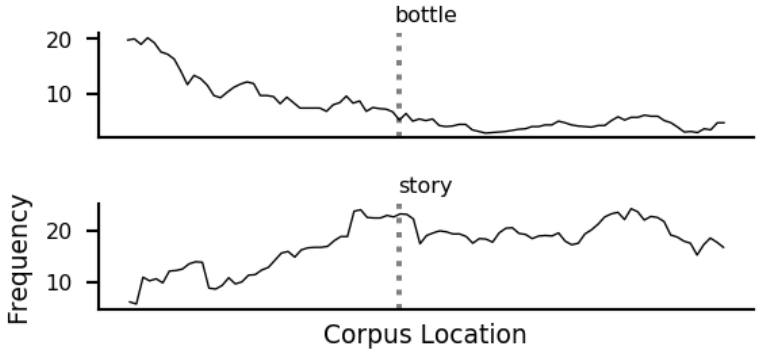

Figure 4.11
Cumulative frequency trajectories of the words bottle and story across 256 partitions of AO-CHILDES.

CHAPTER 5: A GOOD START

In this chapter, I test whether the age-order effect, described in chapter 3 is caused by improper tuning of the SRN's hyperparameters. I show that this is not the case, and conclude that the age-order effect is a psychologically relevant phenomenon rather than an uninteresting detail of how neural networks are trained or tuned. I explore three hypotheses to explain the age-order effect, and settle on the most promising: the 'good-start' hypothesis.

Is the age order effect psychologically relevant?

In the previous chapter, I demonstrated that probe words are not equally distributed across the two partitions of AO-CHILDES. Not only did I find that there are more probe words in partition 1, but the types of probes that occur in one partition tend to occur more frequently in the same partition than in the other. This means that in order to achieve best semantic categorization performance, the SRN must incorporate information contained in both partitions[9]. In other words, the SRN cannot afford to wait; due to the incremental training regime, information about semantic category membership not acquired during training on the first partition may never be seen again by the model. This risk is mitigated by a training regime in which presentation of examples is randomized (there would be no partitions). Because presentation of examples is not randomized, the SRN must ensure that any information about semantic category membership is utilized to the fullest extent possible. But this presents a dilemma, unique to incremental training regimes: The differences in complexity and word

[9] In offline analyses I showed that maximal performance on syntactic category performance does not require integrating syntactic information present in both partitions. Instead, syntactic categorization performance peaks during training on partition 2, and additional training on partition 1 does not improve it.

distributions (e.g. more nouns in partition 1, more verbs in partition 2, etc.) may require that different strategies are used for the acquisition of semantic category knowledge. The same strategy may not suffice for extracting the maximum amount of information about semantic category membership in both partitions. It may be more effective to use two different strategies, to deal, for example, with the fact that semantic dependencies in partition 2 span greater distances or that there is a greater number of conjunctions and prepositions in partition 2.

How does this help us understand the age-order effect? If it is true that the two partitions of AO-CHILDES require different 'strategies' for extracting information about semantic category membership of probe words, then the age-order effect should not be considered a psychologically relevant phenomenon, but an artefact of improper model tuning. A well tuned model (and training regime) ensures that the model can extract the maximal amount of desired information from each training example (Buchnik, 2019). This is a best-case scenario. In most situations, however, a model may see a large number of training examples for which it was not 'maximally ready' for. Most neural network practitioners combat this potential slowdown in learning by presenting training examples in randomized order. The hope is that the model will eventually learn the target function, even if there existed an (unknown, most likely) order of presentation of training examples which would have led to the same level of performance faster. With an incremental training regime, it is less clear how to combat reduced 'model readiness'. Instead of randomizing the presentation of training examples, a practitioner might tune the learning rate decay, momentum, or the optimization method used to update model parameters. Put differently, a practitioner must match the state of the model to the kinds of training examples that the model is exposed to at each stage of training. It is possible that the age-order effect is simply a result of reusing the same hyper parameters in both training conditions. If I had more

carefully tuned the hyper parameters for the SRN trained in reverse age-order, perhaps it would achieve the same level of performance as the SRN trained in age-order. Because I do not think this is actually the case, I will refer to this explanation of the age-order effect as the null hypothesis. Instead, I think that the age-order effect is revealing a more fundamental truth about learning in neural networks (and by extension, humans) that cannot be explained by improper tuning. The goal of this chapter is to test whether the age-order effect is indeed a mere artefact of improper tuning, and ultimately to reject this notion.

My overall goal in this work is not only to explain the age-order effect, but to establish it as an important phenomenon worth studying. To do so, I have to demonstrate that it is not simply an artefact of the intricacies of neural network training and tuning. Instead, I wish to demonstrate that it is a window into deeper questions about the incremental nature of learning in general, that includes both neural networks and humans. I must demonstrate that there exists no 'perfectly tuned' set of hyperparameters which can result in either better or identical performance for SRNs trained in reverse age-order compared to SRNs trained in age-order. If the age-order effect is indeed relevant to theories of learning in general, then a benefit for age-ordered training must be robust against variation in hyperparameters. In no circumstance, should performance in the reverse age-ordered training regime achieve the best possible performance on the semantic categorization task. If this were the case, then my results in chapter 3 could be interpreted as a statistical anomaly resulting from improper exploration of the hyperparameter space.

Tuning the number of iterations per partition

While it is impossible to demonstrate that the age-order effect is robust to all possible hyperparameters, I will demonstrate that it is robust to a hyperparameter I think is the most

relevant. In this section I will explain which hyperparameter I have decided to test and how I have arrived at this decision. I started by looking at the corpus analyses of AO-CHILDES described in the previous chapter. I considered the greater syntactic complexity of partition 2 compared to partition 1. The reduced complexity of partition 1 may make the SRN 'ready' for acquiring semantic dependencies *earlier* compared to the SRN trained on partition 2 first, where greater complexity may delay 'readiness'. Thus, a difference in semantic categorization performance at the end of training may simply be a matter of *when* a model was first 'ready' to acquire semantic dependencies. For example, it is possible that before being able to acquire semantic dependencies, the model trained in reverse age-order, must first learn complex syntactic dependencies which are more abundant in partition 2. This means that by the time the model trained in reverse age-order can begin acquisition of semantic dependencies, the model trained in age-order has already begun the process. In other words, the age-ordered model has a head-start. It is equally possible that acquisition of semantic dependencies during training on partition 2 is simply *slower*, because of the greater complexity in partition 2. For example, semantic dependencies may span greater distances, or may me embedded in more complex or rare constructions in partition 2 than in partition 1.

In sum, the null hypothesis claims that the age-order effect is due to an imbalance (between partitions) in the amount of *time* it takes to acquire semantic dependencies. Underlying the null hypothesis is the assumption that the age-order effect is due to a difference in the *complexity* of the two partitions, rather than the *order* in which that complexity is experienced. If true, the age-*order* effect would have to be renamed to something like the *partition*-effect or *complexity*-effect. A good (though possibly not the best) way to control for the effect that partition complexity might have on the speed of acquisition, is to tune the number of iterations

over each partition. Spending proportionately more time training on partition 2 to partition 1 should counteract the late start or the reduced speed at which semantic dependencies are acquired by the SRN trained in reverse age-order because the SRN would have more time to extract the maximum amount of information about semantic category membership from partition 2. If the age-order effect persists, even after controlling for the time needed to acquire the maximum amount of information about semantic category membership in this fashion, I would conclude that the age-order effect is not an artefact of improper tuning.

I trained at least 3 SRNs on AO-CHILDES either in age-order or reverse age-order, and in one of two tuning conditions: The model either iterates over the first encountered partition 10 times and the second encountered partition 30 times (10-30), or the model iterates over the first encountered partition 30 times and the second encountered partition 10 times (30-10). To be precise, the first number in the label X-X corresponds to the number of iterations over partition 1 for a model trained in age-order, and the number of iterations over partition 2 for a model trained in reverse age-order. If it is true that more training is necessary to extract the maximum amount of semantic category membership information from partition 2 compared to partition 1, then the tuning condition 30-10 should improve balanced accuracy for SRNs trained in reverse age-order, and eliminate any performance difference between the two training regimes at the end of training (the hallmark of the age-order effect). The results, shown in figure 5.1 show that this is not the case. While the tuning condition 30-10 did slightly improve end-of-training semantic categorization performance of SRNs trained in reverse age-order (barely noticeable difference between blue and green line in the right panel), this minimal performance improvement did not eliminate the performance gap between the best end-of-training performance of the SRNs trained in age-order (grey line in left panel) and the best end-of-training performance of the SRNs

trained in reverse age-order (grey dotted line in right panel). Additional simulations with different number of tuning conditions (e.g. 35-5, 5-35) gave the same pattern of results: In each case, the age-order effect persisted. As mentioned before, it is possible that hyper parameters other than the number of iterations per partition may need to be tuned to eliminate the age-order effect. However, I have tested several other hyper parameters (mini batch size, learning rate, hidden size, number of backpropagation-through-time steps, vocabulary size) and in all cases the best performing model was the one that was trained in age-order, beating the best model trained in reverse age order by a clearly noticeable margin. A simple explanation invoking improper tuning of hyper parameters to the different demands of each partition (e.g. 'increased complexity in partition 2 requires more time spent training on partition 2') is therefore ruled out.

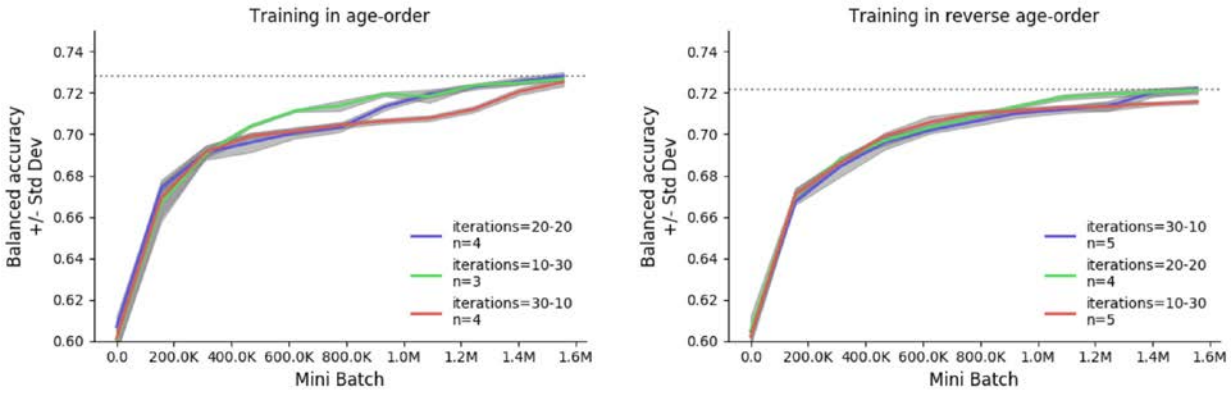

Figure 5.1
Average balanced accuracy as a function of training time (in minibatches) for SRNs trained in age-order (left panel) and reverse age-order (right panel). Color assignment is contingent on performance: For example, when training in age-order, the best performing group of SRNs was trained in tuning condition 20-20, the second best group was trained in tuning condition 10-30 and the third best group was trained in tuning condition 30-10. The best performing group of SRNs trained in reverse age-order however, was trained in tuning condition 30-10.

Quantity vs. Quality

The null hypothesis has been set aside, but what are the alternatives? In fact, it would have been simpler to explain the age-order effect if the null hypothesis had withstood testing. Because this is not the case, an alternative, and probably more complicated, explanation is needed. Instead of explaining 'model readiness' simply in terms of *characteristics of the model* (e.g. number of iterations), 'model readiness' may be a more complex construct, like the *combination of characteristics of the model and the input it has received*. Underlying the null hypothesis is the assumption that 'model readiness' can be varied at any stage during training, regardless of what input the model has already seen or the order in which the input is presented. For example, training on input with reduced complexity first (partition 1) may increase 'model readiness' to acquire semantic dependencies in partition 2. Without having first trained on partition 1, the model trained in reverse age-ordered cannot establish such 'readiness' and semantic categorization performance during training on partition 2 must remain limited, no matter the number of iterations over partition 2. This implies a qualitative difference between partition 1 and 2 in its ability to induce 'model readiness' for acquisition of semantic category knowledge. This qualitative difference cannot be mimicked by simply tuning hyper parameters. Hyper parameters specify how the *model* is to behave during training, but they cannot change the quality of the *input*. Even after controlling for potential differences in the time needed to extract information from each partition, performance is still greater when training in age-order. The lesson is that additional experience with the same material does not guarantee maximal performance. It is not the *quantity* of previous training experiences that matters (e.g. number of iterations), but the *quality*.

Three hypotheses

Given the results of the tuning simulations, it appears the current hyper parameter configuration is already very close to the 'sweet spot'. While varying the number of iterations may increase the quantity of experiences during training on partition 2, in order to achieve maximum performance, it appears that its quality must be improved. No amount of tuning can achieve this. Therefore, partition 1 has some special quality (e.g. reduced syntactic complexity) that partition 2 has less of, and that this quality improves semantic categorization performance only when partition 1 is trained on *first*. This notion implies that the age-order effect is an *interaction between the changing learning dynamics of the model over time and the changing quality of the input over time*. In the space of hypotheses that explain the age-order effect in terms of this interaction, there are three important distinctions to be made. The choice is between what I will refer to as the entrenchment hypothesis, the good-start hypothesis, and the scaffolding hypothesis. In the remainder of this chapter I will describe the three hypotheses, provide some evidence (and counter-evidence) where possible, and explain why I have chosen to pursue the good-start hypothesis over its competitors. All three hypotheses share the notion that partition 1 has a special quality that partition 2 has less of. That quality may be its reduced complexity, or its greater number of probe word occurrences, or something else which I did not detect in my corpus analyses. Whatever it is, performance is boosted during early training in age-order compared to training in reverse age-order, where no such boost occurs. Where the three hypotheses differ, is in explaining how the performance improvement of the model trained in age-order persists until the end of training.

The entrenchment hypothesis

Underlying the entrenchment hypothesis is the assumption that the age-order effect can be broken down into two independently caused effects. The first is the performance improvement observed during the first half of training (early component), and the second is the performance improvement observed at the end of training (late component). See figure 5.2 for a visual breakdown. Under the entrenchment hypothesis, the early component of the age-order effect is driven by the special quality of partition 1. Like the other two hypotheses, the entrenchment hypothesis remains agnostic about what that quality may be. The late component, the greater performance at the end of training in age-order, is caused by the reduced ability of the model trained in reverse age-order to benefit from the performance boost once it has reached partition 1 during the second half of training. Why should the model trained in reverse age-order not benefit equally from the performance boost provided by partition 1? The answer to this question is the core argument underlying the entrenchment hypothesis. The same boost in performance is not achieved because of a gradual reduction of the model's ability to learn new information. While a small boost might occur, it is not enough to result in the same level of performance that the age-ordered model achieves at the end of training. The reduced ability of the model to learn as a function of training time, is often referred to as weight entrenchment (Zevin & Seidenberg, 2002). Weight entrenchment occurs when the error signal used to update a parameter must be back-propagated through a nonlinear activation function (e.g. sigmoid). As the absolute value of the parameters (weights) of the model increase with training, the error signal is pushed closer to zero. The idea is similar to the notion of 'plasticity' in biology, which refers to the magnitude of the effect that a new learning experience can have on the learner. One can think of weight entrenchment as gradually reducing the system's plasticity.

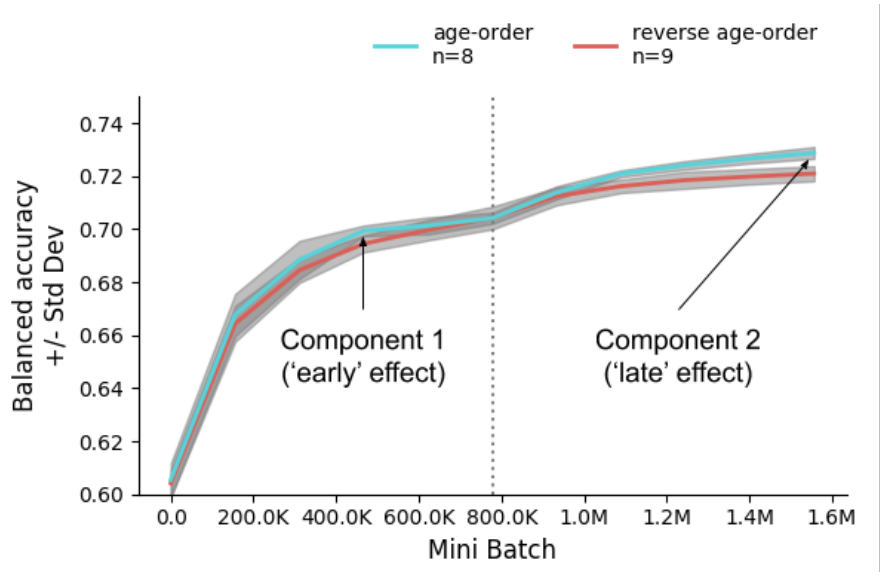

Figure 5.2
A reproduction of figure 2.3 with component 1 and 2 of the age-order effect indicated.

While I do not actually subscribe to the entrenchment hypothesis, it will be informative to explore the evidence that is consistent with it. As we shall see, there is very little support for this idea. This has likely more to do with the difficulty of measuring entrenchment. After all, there are no standard and straightforward methods for investigating the internal representations learned by the SRN, let alone how quickly they are modified. Thus we must keep in mind that the lack of evidence is not evidence of absence. One straightforward way to support the entrenchment theory is not by directly measuring weight entrenchment, but by demonstrating that the greater performance of the model trained in age-order is not due to a performance *improvement* per se, but due indirectly to a performance *reduction* of the model trained in age-order. It could be said that training in age-order doesn't actually provide any advantage (e.g. relative to a model trained in random order); the advantage only appears when it is compared to a model trained in reverse age-order. To test this idea, I broke AO-CHILDES into 256 equally sized partitions. This

allowed me to train a model on randomly ordered input, where the order of the 256 partitions was shuffled. Additionally, I trained SRNs on partitions in age-order (no shuffling) and in reverse age-order (no shuffling, but reversing of the order), to compare their performance to the model trained in random order. I found that the models trained in reverse age-order achieves worse performance than the models trained in random order. To be precise, semantic categorization performance of the models trained in random order is halfway between the performance of models trained in age-order and reverse age-order, at all but the earliest and latest evaluation time points[10]. This observation is consistent with the idea that weight entrenchment is *reducing* the overall performance of a model trained in reverse age-order. A model trained on examples in random order would at least see *some* examples of partition 1 during early training, at which weight entrenchment has not yet reduced the model's ability to benefit from the performance boost provided by those examples. Because the model trained in reverse age-order sees *none* of the examples in partition 1, performance is reduced.

Additional observations consistent with weight entrenchment comes from an analysis in which I re-initialized a randomly chosen 90% of recurrent weights ten times during training at equally spaced intervals (after evaluating the balanced accuracy). This combats weight entrenchment because the level of entrenchment of a weight is proportional to the weight's distance from zero. The larger its absolute value, the less responsive it is to subsequent updates. Re-initialization simply sets weights close to zero. (Re-initialization is identical to initialization of weights before the start of training, except that only a random 90% of the recurrent weights are affected.) The results of raining with weight re-initialization are shown in figure 5.3. Panels are ordered left to right from indicating the proportion of recurrent weights that are re-initialized

[10] With 256 partitions, the improvement of the models trained in age-order did not last until the end of training. I will discuss this finding in a subsequent section.

(10%, 50% and 90%). Re-initializing 90% of the recurrent weights during training causes the age-order effect to shrink considerably. This could be interpreted in many different ways, but this analysis was specifically designed to test whether a reduction in weight entrenchment can reduce the age-order effect. Because the age-order effect was reduced, one might conclude that weight entrenchment has an important role to play in explaining the age-order effect.

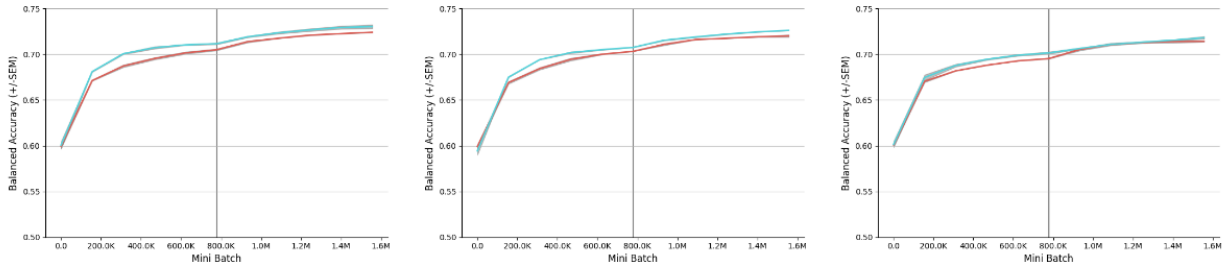

Figure 5.3
Average balanced accuracy as a function of training time (in minibatches) for SRNs trained in age-order (blue line) and reverse age-order (red line). A randomly chosen 10% (left panel), 50% (middle panel), or 90% (right panel) of the recurrent weights were re-initialized ten times during training, at equally spaced intervals[11].

The scaffolding hypothesis

The entrenchment hypothesis assumes that the age-order effect arises due to training on the most informative portion of the input at the start of training. It does not matter whether the quality of the *input* changes systematically afterwards, because the learning dynamics of the *model* is changing irrespective of the input. Instead, the scaffolding hypothesis proposes that the change in the quality of the input matters just as much as the change in the learning dynamics of the model. In fact, the scaffolding hypothesis requires that they change *simultaneously* to achieve maximal performance. Under the scaffolding hypothesis, partition 1 remains special, relative to

[11] Unrelated to the discussion in this chapter, these results show that semantic categorization performance is almost immune to re-initialization of recurrent weights. This shows that most of the knowledge about semantic category membership is not stored in the recurrent weights of the SRN, but in the input and/or output weights.

partition 2, in terms of providing greater performance relative to partition 2, but partition 2 takes on an equally important role. It no longer suffices to say that the SRN can only take advantage of the performance boost provided by partition 1 when trained on *at the start of training*, when weight entrenchment is weakest. Instead, the scaffolding hypothesis proposes that it is the *order* of the partitions that accounts for the persistent performance improvement (improvement lasts until the end of training). Under the scaffolding hypothesis, the age-order effect is not broken down into two components (early and late); instead, the improvement that is observed both early and late during training of the age-ordered model are both due to the same cause. The cause is the scaffolding provided by the synergistic change in the structure of the input and the model learning dynamics *over the total duration of training*. Scaffolding guides the descent down the error surface of the model in a manner that is faster, and may arrive in a region associated with lower error. Scaffolding is not just a one-time event, but is a phenomenon that is continuously present and positively influences all subsequent learning experiences. Under the scaffolding hypothesis, the age-order effect is a single phenomenon caused by simultaneous changes in the learning dynamics of the model and the input. Both the model and the input must co-evolve in just 'the right way' to result in enhanced performance at the end of training.

Because no scaffolding occurs when the order of the input is reversed, performance suffers in the reverse age-ordered training condition. But it is important to keep in mind that the scaffolding hypothesis does not specify whether performance of a model trained in reverse age-order is worse compared to a model trained in random order. In both cases, the order of the input is incompatible with scaffolding. The only prediction that the scaffolding hypothesis can make in such a scenario is that the performance of either model should be lower compared to a model trained in age-order. An alternative version of the scaffolding hypothesis might claim that

training in reverse age-order does not result in lower performance simply because training in age-order results in *greater* performance. Instead, it could be argued that training in exactly the opposite direction of the trajectory necessary for scaffolding, results in an *additional reduction* in performance.

The scaffolding hypothesis does not rule out the possibility that weight entrenchment is occurring. Weight entrenchment is a phenomenon that likely affects any simulation where nonlinear networks are trained. While the scaffolding hypothesis accepts that weight entrenchment occurs, it does claim that the cause of reduced performance in the revere age-ordered training condition is not a result of weight entrenchment.

The reader may ask how scaffolding is different from weight entrenchment. In both cases, the model is undergoing some special change during early stages of training which allows it to achieve better performance at the end of training. What distinguishes the two is that weight entrenchment results in reduced acquisition of knowledge, while scaffolding does not. Scaffolding is additive: It acts like a guide, leading the model on a trajectory through parameter space which puts the model in a better position to acquire new (possibly more complex) information. As such, performance rises faster, and may reach a greater performance overall. Weight entrenchment, on the other hand, is subtractive: It does not improve performance directly; rather it does so indirectly by reducing the performance achieved by training a model in reverse age-order. Notice also that the entrenchment hypothesis (and the good-start hypothesis) requires that the age-order effect must be decomposable into two independently caused effects (early and late components). The scaffolding hypothesis has no such requirement. It proposes that the early and late component are both the result of continuously ongoing scaffolding, which influences performance from start to end of training.

An idea that frequently comes to mind when thinking about these three hypotheses is whether catastrophic interference (also known as catastrophic forgetting), should be at the foundation of a fourth hypothesis. Catastrophic interference is a form of forgetting in a neural network that is caused by interference of new knowledge with existing knowledge. It is especially prominent in situations in which the input is non-stationary, meaning some quality of the input is changing over the course of training. In my view, scaffolding is intimately related to catastrophic interference, and does therefore not warrant a fourth hypothesis. Let's consider for a moment what would happen if we considered a fourth hypothesis based on catastrophic interference. Such a hypothesis would assert that the reason that the model trained in reverse age-order does not reach the same level of performance as the model trained in age-order is that it encountered a greater amount of catastrophic interference compared to a model trained in reverse age-order. Presumably catastrophic interference would occur at the partition boundary, and would be greater when transitioning from partition 2 to partition 1 compared to the other way around. But why would there be more interference when training in one order compared to another? One might say that the knowledge that the model acquires during training on partition 2 does not generalize to partition 1 as well as the other way around. This is very plausible. But at this point, the hypothesis is barely distinguishable from the scaffolding hypothesis. As I said before, scaffolding guides the descent down the model's error surface. Like an optimization strategy (e.g. momentum), scaffolding makes this descent more efficient, possibly by resulting in fewer turns, and additionally by starting the descent in a more favorable location. This is equivalent to saying that the 'goal' of scaffolding is to minimize the cumulative amount of interference. Put differently, scaffolding happens when consecutive training examples result in smaller error than randomly ordered training examples. This means that during the course of

training, the total influence of catastrophic interference is reduced. I think that the claim ‘catastrophic interference is greater when training in reverse age-order’ is equivalent to the claim just made, which is ‘catastrophic interference is reduced when training in age-order’.

Support for the scaffolding hypothesis comes from an analysis where I split the balanced accuracy evaluation into two separate measures. One quantifies semantic categorization performance for probe words that occur more frequently in partition 1, and another quantifies semantic categorization performance for probe words that occur more frequently in partition 2. Tracking, during training, these two separate indicators of semantic categorization performance revealed that training on partition 1 first results in an almost equal improvement in both measures, meaning that whatever is learned about the semantic category structure in partition 1 generalizes to probe words that occur more frequently in partition 2. Crucially however, training on partition 2 first did not result in the same pattern; instead, the balanced accuracy for probe words that occur more frequently in partition 1 did not rise as quickly as the balanced accuracy for probe words which occur more frequently in partition 2. This means that the knowledge of semantic category membership acquired during partition 2 does not generalize to partition 1 as well as the other way around. When generalization does not occur, new knowledge must be acquired, and this is can lead to catastrophic forgetting. Crucially, catastrophic forgetting would be more prominent during training in reverse age-order compared to age-order.

There is some computational support for the scaffolding hypothesis. Prior research has shown that presenting an initially restricted hypothesis space, and then gradually expanding it, can be beneficial when learning grammatical categories (Cameron-Faulkner et al., 2003).

The good-start hypothesis

The good-start hypothesis also posits that the initial performance improvement of the model trained in age-order is due to some special quality of partition 1 relative to partition 2. But it does not assert that the age-order effect consists of two independently caused effects. Instead, the improvement during early age-ordered training has the same cause as the improvement at the end of training. Rather than requiring an additional factor, such as weight entrenchment, to explain the improvement at the end of training, the good-start hypothesis proposes that training on partition 1 is the cause of both improvements. This is because training on partition 1 has a lasting effect on performance, which can, and does, last until the end of training. What is the nature of this lasting effect? The good-start hypothesis proposes that training on partition 1 immediately orients the model to a location in parameter space which facilitates subsequent learning. Simply put, the model is provided a 'good start' in parameter space, and this has lasting consequences.

Importantly, the good-start hypothesis remains agnostic about the order of the input; what matters most is where the model *starts* in parameter space. Where it goes does not matter. This means, that the good-start hypothesis doesn't take into consideration the quality of partition 2. Whether or not partition 1 was followed by a partition with equal ability to provide a 'good start' does not matter, because a 'good-start' cannot be provided twice. The scaffolding hypothesis, however, takes into consideration *both* partitions (specifically, their order), rather than considering only what comes *first*. It is in fact difficult to distinguish the good-start hypothesis form the scaffolding hypothesis. One might view the good-start hypothesis as a weaker version of the scaffolding hypothesis if one removed the part that specifies that scaffolding must act *continuously* rather than only at the start of training. Another way to understand the difference is

how each explains the persistence of the performance improvement of the model trained in age-order: Under the scaffolding hypothesis, the persistence of the performance improvements due to the *relationship* of the input seen first with input seen later, while under the good-start hypothesis, the persistent improvement is due to a lasting effect of having trained on partition 1 when the model had not yet seen any training examples.

In my view, the good-start hypothesis provides the best interpretation of the age-order effect so far. In the remainder of the chapter, I will present analyses that support this claim. First, I want to show that the observations described above are equally consistent with the good-start hypothesis. The first observation, that training on 256 reversed age-ordered partitions results in slightly worse performance relative to training on AO-CHILDES on 256 randomly ordered partitions is not something that directly follows from good-start hypothesis, because the good-start hypothesis is strictly about the benefit of training in age-order and does not make predictions about any reduction in performance when training in reverse age-order. But it can accommodate this observation with an additional constraint. We simply need to add the constraint that partition 2 acts as a 'bad-start' which reduces performance by orienting the model to an area in parameter space that impairs, rather than facilitates, subsequent learning. This modification is not so ad-hoc as it might appear, because I have already shown that partition 2 is more complex, and that semantic dependencies are therefore more likely to span greater distances. Learning distant nonadjacent dependencies is notoriously difficult for the SRN, and therefore the kinds of dependencies that the SRN will most likely learn first during training on partition 2 are syntactic in nature. This might actually put the SRN in a worse position to learn semantic categories compared to a randomly chosen mid-section of AO-CHILDES which an intermediate level complexity.

Let's revisit the observation that the balanced accuracy for probes that occur more frequently in partition 1 compared to partition 2 is lower than the balanced accuracy for probes occurring more frequently in partition 2 compared to partition 1 when training on partition 2 first, and that no such difference exists when training on partition 1 first. The interpretation I have given above is that semantic category knowledge acquired during training on partition 2 generalizes less well to partition 1 than the other way around. While this supports the scaffolding hypothesis, which is built on the idea that generalization is not symmetrical across the partition boundary, this finding is also consistent with the good-start hypothesis. This is true because the good-start hypothesis is only a weaker version of the scaffolding hypothesis. In the scaffolding hypothesis, the age-order effect is caused by the combination of seeing partition 1 first and seeing partition 2 last, whereas in the good-start hypothesis the age-order effect is caused by training on partition 1 first, regardless of what comes second as long as the input is drawn from AO-CHILDES (or similar child-directed speech). In the good-start hypothesis, the role that the input plays in the age-order effect has been reduced. But both hypotheses predict that generalization would be better when training in age-order.

The observation that periodic re-initialization reduced the age-order effect is also consistent with the good-start hypothesis. While it is true that reinitialization of weights can be thought of as combating weight entrenchment, it is possible that the *knowledge* encoded in the weights, not their *magnitude* was responsible for the reduction in the age-order effect. Under the entrenchment hypothesis, the age-order effect was reduced because it allowed the models trained in reverse age-order to remain sensitive to new learning experiences and therefore maximally benefit from the performance boost provided by partition 1. But it must be kept in mind that recurrent weight re-initialization not only reduces weight entrenchment, but also erases the

knowledge encoded in the recurrent weights. Thus it is possible that erasing the knowledge also partially erased the 'good start' provided by having trained on partition 1 at the beginning of training. This additional (previously unforeseen) effect of re-initialization would also result in a reduction of the age-order effect.

Evidence in support of the good-start hypothesis

How have I settled on the good-start hypothesis? After all, I have not presented any analyses which clearly dissociate the effects of weight entrenchment, scaffolding, and starting training in a 'good' location in parameter space on semantic categorization performance. In this section, I will explore three lines of evidence in which I have been able to create situations in which these mechanisms should produce different effects

First, there is a straightforward way to dissociate weight entrenchment from scaffolding and a 'good-start'. The analysis is similar to the one in which I trained SRNs on 256 randomly ordered partitions of AO-CHILDES. While the results of that analysis were in agreement of the weight entrenchment hypothesis, a similar situation can be created in which the results violate the entrenchment hypothesis. Instead of breaking AO-CHILDES into 256 partitions and shuffling their order, a similar 'shuffled' training condition can be created, without modifying the 2-partition structure. This can be achieved by populating two equally sized partitions with documents drawn randomly from AO-CHILDES. This was done to keep the simulations in line with previous analyses in which only 2 partitions were used for training. (More about the 2-partition structure vs. the 256-partition structure below.) The motivation for this analysis is to test (again) a core assumption of the entrenchment theory. Entrenchment theory predicts that training in any random order should result in a performance that is both worse compared to

training in age-order and greater compared to training in reverse age-order. The reason is the following: After shuffling the documents, all the information about semantic category membership is evenly spread among both partitions. Performance should be worse in this condition compared to age-ordered training because more semantic category information occurs in the second half of training, where entrenchment would reduce the ability of the model to acquire it. Similarly, performance should be better compared to training in reverse age-order because more semantic category information is seen during the first half of training when acquisition of new knowledge is most effective. The results of both simulations are shown in figure 5.4. The right panel shows performance in the three conditions when training iterates over 256 partitions. The results are in agreement with the entrenchment hypothesis which predicts that performance achieved in the 'shuffled' condition should be intermediate between performance achieved by models trained in age-order and models trained in reverse age-order. The left panel shows an entirely different story, however. When the 2-partition structure is preserved, the performance in the 'shuffled' condition is well above that of the other two conditions. This is strong evidence against the entrenchment hypothesis, which predicts that performance in the 'shuffled' condition should be intermediate, regardless of the number of partitions.

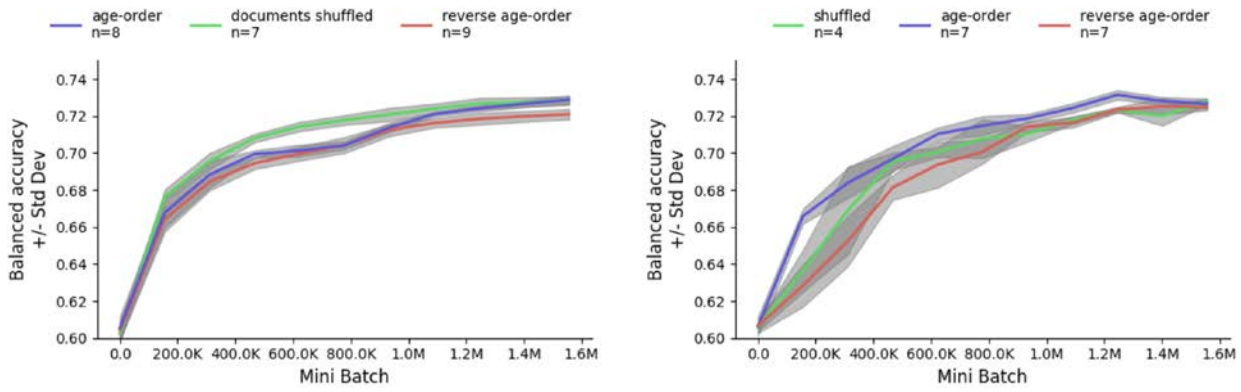

Figure 5.4
Average balanced accuracy as a function of training time (in minibatches) for SRNs trained on 2 partitions (left panel) and 256 partitions (right panel). The color of the line indicates the training order. The pattern of performance in the 'shuffled condition (green lines) varies with the number of partitions.

It's worth pausing for a moment to think about the dramatic departure in the results shown in the left panel compared to those shown in the right panel. What are some possible explanations for this large improvement in performance in the 'shuffled' condition when the 2-partition structure is preserved? First, while performance is well above performance achieved in the other two conditions during the first half of training, this improvement does not last until the end of training. In fact, performance is indistinguishable between models trained in age-order and models trained in random order. This is important because the goal of this work is to support the idea that age-ordered training can result in a consistent performance improvement, and is not just an artefact due to statistical sampling or some psychologically irrelevant component of the training conditions. End-of-training performance achieved by a model trained in age-order should at least be better than a model trained in some baseline condition (e.g. random order) for it to be considered relevant to learning theorists. What about the large improvement during early training? This is most likely caused by increasing the diversity of the training examples that are in each partition. Due to random assignment of documents to one of two partitions, each partition

contains a maximally diverse sample of AO-CHILDES. This means that the model in the 'shuffled' training condition sees examples at both extreme ends of the complexity dimension (and any other dimension along with training examples vary). It is well known that neural network training benefits from training on input that is most representative of the input as a whole. The reason that performance is greater than even that of the models trained in age-order is that he models trained in age-order experience a very non-diverse set of training examples. For example, partition 1 of AO-CHILDES contains the least complex utterances compared to the average complexity of AO-CHILDES as a whole. The same is true when training reverse age-order; the training examples are skewed towards the higher end of the complexity continuum, and thereby represent relatively unrepresentative examples of the input as a whole. The model that encounters a representative sample of the data right away is able to more faithfully represent the structure of the input. Doesn't this interpretation violate the assumption of the good-start hypothesis, which claims that starting training on partition 1 (which is not representative of the input as a whole) results in better performance than starting training on some randomly chosen section of the input? After all, the models trained in random and age-order achieve the same end-of-training performance. It's hard to say, because the performance achieved in the two conditions may be at ceiling. This means that the performance of models trained in age-order may have been larger, had there been room for improvement. This is not a satisfying response, and is worth further investigation.

Something I have glossed over is that the improvement in performance of the age-ordered model trained on 256 partitions does not last until the end of training. There is clearly a strong early boost in performance in the age-ordered training condition, but this improvement is not persistent. The reason the age-order effect is so intriguing is because of the performance

difference at the very end of training. So why was it not observed here? Under the good-start hypothesis, this result is actually not surprising at all. The good-start hypothesis explains the persistent improvement in terms of having started training in a good location in parameter space, which helps to orient future training experiences. However, the long-lasting benefit of a 'good start' cannot last forever. When there are only 2 partitions, the effect need not last long; it only needs to persist for the second half of training to be able to affect end-of-training performance. When there are 256 partitions, however, the 'good start' provided by the earliest partitions must last much longer in order to affect end-of-training performance. The good start provided by the first partition must endure training on 255 subsequent partitions. When training over 2 partitions only, the last weight update in partition 1 is only 1 half (505) of the total number of weight updates removed from the very last weight update at the end of training. When there are 256 partitions, however, the last weight update in partition 1 is 255 256^{ths} (99%) of the total number of weight updates removed from the end of training. Under the good-start hypothesis, the models trained in this condition had a 'good start' but the benefits were not able to withstand the greater number of weight updates that occurred between the 'good start' and the end of training.

There is more to be said about the lack of persistent performance improvement when AO-CHILDES is broken into 256 partitions. For example, what would the scaffolding hypothesis predict in such a scenario? Because scaffolding is continuously acting on the model during training, the same defense used above cannot be used. Moreover, because the consecutive age-structure is preserved when training in age-order, the scaffolding hypothesis should predict an age-order effect accompanied by greater performance at the very end of training. In fact, the scaffolding hypothesis predicts a *larger* age-order effect (including a larger performance improvement at the end of training) when training in age-order on 256 partitions compared to

training in age-order on 2 partitions because the age-structure is more pronounced when there are a greater number of partitions. The only age-structure a model has access to is the order relationship *between* partitions; any age structure *within* partitions is lost due to iterating multiple times over each partition. A consequence of having more partitions is therefore a greater preservation of the age-structure. But because no end-of-training performance improvement was found, let alone a greater performance improvement compared to training on 2 partitions, a basic principle of the scaffolding hypothesis is violated. It is true that the performance improvement during the first half of training is greater when AO-CHILDES is broken into 256 partitions compared to 2 partitions. But, the scaffolding hypothesis also asserts that if there is an early improvement when training in age-order, there should also be an improvement at the end of training. This is because the scaffolding hypothesis does not consider an early improvement to be caused independently from an end-of-training improvement. If there is one, there must be the other. Figure 6.3 clearly demonstrates that an early improvement *can* occur without an end-of-training improvement (256 partitions, right panel). By the same logic, however, this would appear to invalidate the good-start hypothesis which also claims that the age-order effect is due to a single cause, and not two. But not so fast; I have already explained why this pattern can still be considered consistent under the good-start hypothesis without violating any of its core principles. An early improvement (as observed in the right panel of figure 6.3) *may* be followed by a late improvement, but crucially, it is not *required*. The good-start hypothesis only specifies that if a late improvement did occur, it would be an (optional) effect of the same factor that caused the early improvement. The end-of-training performance improvement is optional because there are conditions under which a 'good start' is too distant from the end of training (in number of weight updates) to influence end-of-training performance. One of these conditions, as

we have seen, is training on a larger number of partitions. Considering all of the above, the good-start hypothesis has, in my view, more merit than its competitors.

Additional evidence in support of the good-start hypothesis comes from training on 4 partitions of AO-CHILDES, in which the two middle partitions are seen during the first half of training, in both training conditions (age-ordered vs. reverse age-ordered). Because both the good-start and the scaffolding hypotheses assert that the age-order effect can only occur when training starts on a relatively early section of AO-CHILDES (e.g the age of the target child is below average), these two hypotheses predict that the age-order effect does not occur. Keep in mind that the scaffolding hypothesis requires both the age-structure of the input to be preserved during training, *and* a 'good start' provided by training on an early section of AO-CHILDES. Because in this simulation training starts on a mid-section (the middle half of AO-CHILDES), the latter requirement does not hold. What would the entrenchment hypothesis predict? If it is true that the reduced performance of the model trained in reverse age-order is a consequence of weight entrenchment (which reduces its ability to learn valuable information during late-stage training on an early section of AO-CHILDES), then training on the middle half of AO-CHILDES first, in both conditions, should reduce the age-order effect. This should occur because the original partition 1 is no longer seen *first* during age-ordered training, when the model is supposed to be most able to acquire new knowledge. But, critically, the age-order effect should not disappear entirely, because a portion of the original partition 1 is still seen *earlier* compared t*o* training in reverse age-order. The age-order effect should still occur, in contrast to the other two hypotheses, because the entrenchment hypothesis does not require a 'good start'. As long as the portions of the original partition 1 (which provides a boost in performance relative to the original partition 2) is seen *earlier*, then performance should be higher. This setup

therefore allows us to distinguish the entrenchment hypothesis from the good-start and the scaffolding hypothesis. The results are shown in figure 5. 5. There is no difference in performance between the two training conditions during the first half of training, as expected. The blue line then rises above the red line, indicating slightly improved performance as a result of training on a subsection of the original partition 1 (the first half) compared to a subsection of the original partition 2 (the last half). This is in line with all previous observations that training on the original partition 1 provides a boost in performance compared to training on a latter section of AO-CHILDES. The most important finding is what happens next; performance quickly flips when the models trained in the reverse age-order condition now train on a subsection of the original partition 1, and the models trained in age-order now train on a subsection of the original partition 2. The result is a sort of anti-age-order effect, because the performance of the models trained in reverse-age order (at least during the second of training) is greater, at the end of training, compared to models trained in age-order. This is the opposite of what should have occurred under the entrenchment hypothesis, which predicted an age-order effect (albeit smaller than previously observed). The intuitive explanation why the entrenchment hypothesis fails here is that the models trained in reverse age-order should not have benefited as much as they did during training on a subsection of the original partition 1 at such a late stage of training. Weight entrenchment, by then, should have severely reduced the ability of the models to benefit from training on a subsection of the original partition 1. In fact, the reversal, clearly visible at the 1.2 million mini-batch mark, is strong evidence in favor of the idea that the SRN, as implemented in this work, is very capable of acquiring new knowledge, even after 1.2 million weight updates.

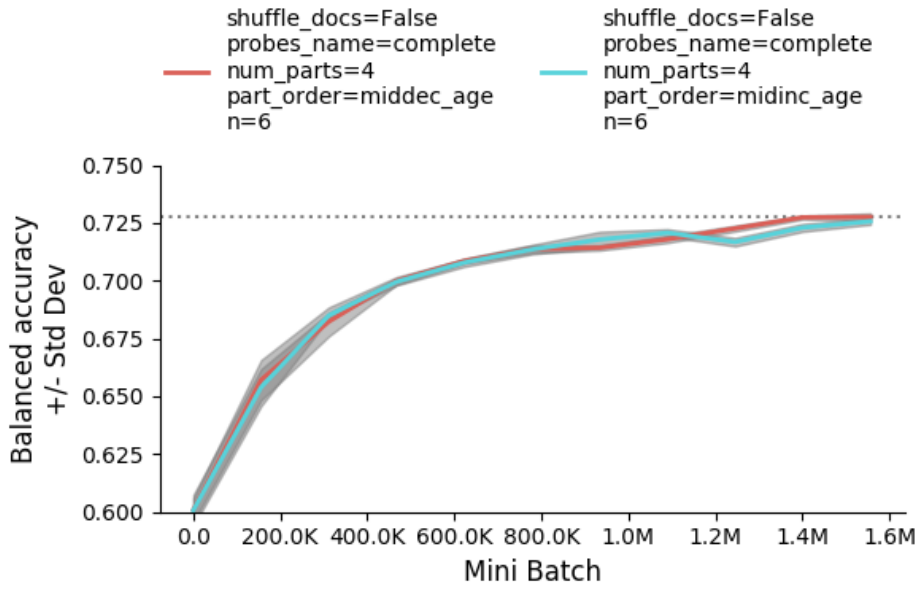

Figure 5.5
Average balanced accuracy as a function of training time (in minibatches) for SRNs trained on 4 partitions ordered by age of the target child (blue line) and in reverse (red line). In both conditions, SRNs were trained on the two middle partitions first. The remainder of the training regime is unchanged, save for the absence of the middle partition.

A final observation to counter the entrenchment hypothesis makes use of the prediction error across training. During training, in both conditions, I tracked the perplexity associated with each mini-batch, which is simply the exponentiation of the cross-entropy loss (used to quantify prediction error and to train the model) with Euler's number as the base (see chapter 2). In both training conditions, the perplexity should gradually reduce as a model is acquiring more information about the sequential structure of AO-CHILDES. I should also be able to detect a greater drop in perplexity when comparing the first half of training between a model trained in age-order and a model trained in reverse age-order. Partition 1 is less complex (see chapter 4), and this should make sequence prediction easier. The question I am interested in, however, is whether this trend will hold when comparing the second half of training. Under the entrenchment hypothesis, this trend should not hold, because weight entrenchment has reduced the ability of

models in both training conditions to acquire new knowledge. Rather than seeing the same trend in the first half of training mirrored (symmetrically) in the second half of training, the entrenchment hypothesis predicts that perplexity should continue to be lower (indicating better sequence prediction performance) for a model trained in age-order compared to a model trained in reverse age-order, despite the fact that the input to the two models have switched. If weight entrenchment was not reducing either models' ability to acquire new knowledge during the second half of the training, the model trained on partition 1 during the second half of training should achieve performance at least as good as that achieved by the model trained on partition 1 first (the model trained in age-order). In other words, the perplexity curves should flip at the partition boundary, in a relatively symmetrical fashion. Any departure from this symmetry would indicate that weight entrenchment (or a similar mechanism) has reduced the ability of both models to acquire new knowledge. The results of this analysis are shown in figure 5.6. As expected, I observed a prominent spike in perplexity at the partition boundary in both training conditions. This is evidence that the knowledge about the sequential structure of one partition does not generalize perfectly to the other transition. However, both models are able to recover from the transient increase in error. In fact, their recovery is so strong, that the performance flips. During training on partition 1, models trained on partition 2 first are able to achieve perplexity as low as that achieved by models trained on partition 1 first. Thus, perplexity at any point in time does not appear to be influenced by *how long ago* partition 1 has been trained on, but rather whether partition 1 is *currently* trained on. This symmetrical pattern of performance is strong evidence against the entrenchment hypothesis.

While this analysis can clearly dissociate entrenchment from scaffolding or a 'good start', it cannot differentiate the latter two. One of the reasons is that that perplexity is an error signal

associated with a model's success at sequence prediction, and not related to the model's knowledge of semantic categories. It measures how well a model can predict *any* next word; it is not limited to the case where *only* semantic dependencies are predicted.

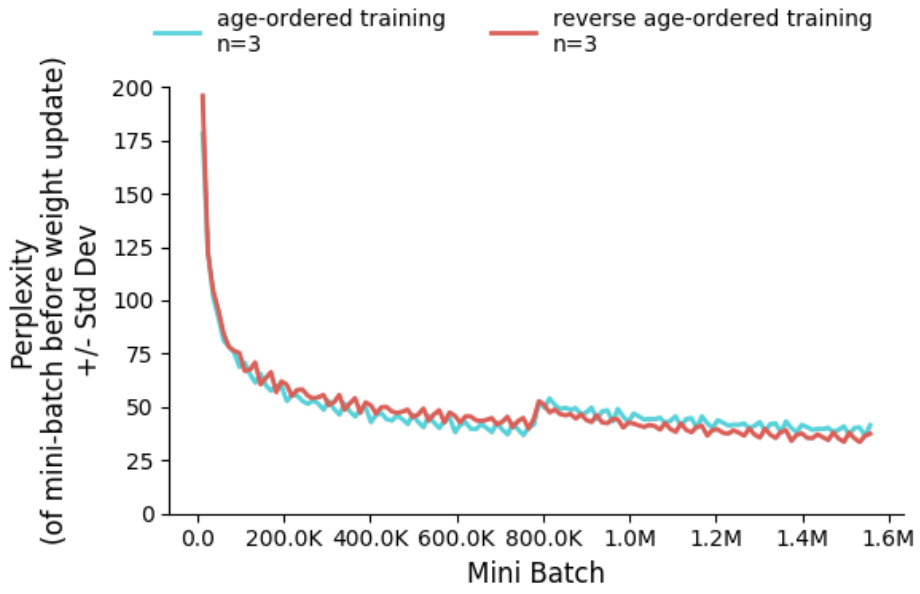

Figure 5.6
Average perplexity as a function of training time (in minibatches) for SRNs trained in age-order (blue line) and in reverse age-order(red line). Perplexity was calculated as ana average over each sequence in a mini-batch, before weight updates associated with the mini-batch have been applied.

When incremental training is useful

I have shown in figure 6.3 that performance of models trained on 2 partitions which are composed of random samples of AO-CHILDES is as good as that achieved by models trained in age-order. This reduces the credibility of the starting-good hypothesis, because it implies that starting on partition 1 did not provide any performance benefit compared to starting training on a partition with randomly chosen examples. I already said that this may have to do with the possibility that performance is at ceiling, meaning no further improvement can be made.

Disregarding this possibility, this finding constrains the conditions under which partition 1 is actually a good section of AO-CHILDES to training on first. When nothing is known about the structure of the input, the best 'starting' partition is the *whole* input, and not some subsection. There is, after all, a good reason why neural networks are not typically trained incrementally. Splitting the input into partitions and training incrementally over the partitions such that no partition is ever revisited (as is done here), typically results in worse performance relative to training on the input as a whole[12]. This is because the incrementally trained model can only integrate over information contained in one partition, rather than over the entire input. At any point during training, the knowledge encoded in its weights is either more heavily representative of one or the other partition, and not their combination. This is a warning to readers who are tempted to try incremental training regimes; without a good reason to break training apart, incremental training is not advisable. This leads me to an important question: Why investigate incremental training regimes when non-incremental training is more likely to produce better results? The answer has to do with psychological plausibility. A child experiences language sentence by sentence, and conversation by conversation; it is unlikely that integration takes place over experiences lasting longer than a day (e.g. sleep consolidation). Only in such cases, when the input is processed in an online or close-to-online fashion, does it make sense to think about the incremental structure of the input. In such a case, the goal is no longer to achieve the *best* possible performance, but to minimize the *reduction* in performance due to not having access to all of the input (e.g. a lifetime worth of language experiences).

[12] I found this to be true even with AO-CHILDES. Non-incremental training (iterating 20 times over the entire input) results in slightly better performance than iterating 20 times over partition 1 and 20 times over partition 2 ('training in age-order').

With this in mind, it is worth re-examining the good-start hypothesis. The results of figure 5.3 suggest that partition 1 is not the *best* starting partition. It is probably true that it is better than *partition 2*, but is it better than a *random sample* of AO-CHILDES? The results shown in figure 5.3 demonstrate that the answer is no. But, returning to the hypothetical child who experiences language in presumably much smaller partitions than 2.5 million words (roughly the size of one partition), the question of what input to start training on is much more important. Experiencing language in 256 partitions, as explained in chapter 2 is psychologically plausible, and it is in this scenario where age-ordered training shines. Now, starting training on a randomly chosen section of AO-CHILDES results in worse performance (at early stages during training) compared to starting on the first of 256 age-ordered partitions. The trend is reversed the size of the partitions is changed from very large to a size that is psychologically plausible. This means that the lessons learned in this work are more relevant to the challenges faced by children who must make the most of their incrementally structured experiences, rather than to neural network practitioners hoping to improve performance of their models. Put differently, the study of the age-order effect is not about how to improve performance, but how to protect from a loss of performance due to the constrained learning environment faced by children. When the learning environment is made less incremental, the benefit of age-ordered training (or any other kind of incremental training regime) is reduced, and - at the limit - can become detrimental to performance.

Limitations

Because I am not using the term scaffolding to refer to a specific concept, the notion of scaffolding is vague and not well developed in this chapter. I tried to relate scaffolding to

interference, when I said that scaffolding occurs when consecutive training examples result in smaller errors than randomly ordered training examples. But we must keep in mind that the goal is to acquire knowledge about the semantic category structure AO-CHILDES and not the sequential structure of AO-CHILDES. While the latter is required to do the former, at the limit, better performance on sequence prediction may harm semantic category acquisition. The problem is that training error is only indirectly related to semantic category knowledge. The model should be able to predict words relatively well in order to obtain an understanding of the semantic category structure, but if the model starts to memorize the training examples, item by item, then this would impair semantic category knowledge. In sum, scaffolding remains ill-defined, and needs to be grounded quantitatively, but not solely in a measure of sequence prediction error.

Another issue concerns the portrayal of the entrenchment theory. I characterized weight entrenchment as a static property of a model, which is proportional to the number of cumulative weight updates. But it is possible that weight entrenchment occurs in proportion to the kind of input trained on. It might occur sooner when the input is more complex or less regular. This means that partition 1 or partition 2 might induce weight entrenchment faster than the other. If true, the findings described above to argue against the entrenchment hypothesis would be largely invalidated. Moreover, this would blur the line between the entrenchment hypothesis and the other two hypotheses and make them even more difficult to tease apart experimentally.

Another explanation of the age-order effect that has been largely ignored is the possibility that the knowledge about semantic categories acquired during partition 2 is more likely to be forgotten during training on partition 1 compared to the other way around. If so, partition 1 would have to be re-characterized as a ‘bad end’ rather than a ‘good start’. In fact, this kind of

explanation is consistent with the analysis reported in chapter 3 in which it was found that the performance improvement at the end of training of the model trained in age-order is largely restricted to probe words that tend to occur more frequently in partition 2 compared to partition 1.

CHAPTER 6: A THEORETICAL FOOTHOLD

In this chapter, I outline a basic theoretical understanding of the age-order effect that is built on the good-start hypothesis developed and defended in the previous chapter. First, I develop a more detailed understanding of the hypothesis, both in terms of the input, and the model. Then, I provide some initial support for the theory from simulations I have already conducted. Lastly, I generate predictions for several experiments, the results of which are reported in chapter 7.

What is a good start?

The goal of this section is to address the question of what exactly it means to 'start good'. The word 'good' simply refers to a 'good' level of performance achieved by the SRN. Crucially, performance does not mean *sequence prediction* performance. I have conceptualized the good-start hypothesis entirely in terms of *semantic categorization* performance. Henceforth, I will refer to semantic categorization performance simply as 'performance', unless explicitly told otherwise (e.g sequence prediction performance). Already, one might wonder whether this is a mistake, and that it would be better to develop a theory based on what the SRN is actually trained to do. I agree that a satisfying explanation of the age-order effect ultimately requires a description at the level of the sequence prediction machinery. However, I have found sequence prediction to be a bad measure of semantic categorization performance. In fact, at small timescales, prediction performance does not detectably covary with semantic categorization performance. Even conceptually, It is not at all clear that an improvement in semantic category knowledge must always be accompanied by better sequence prediction performance. The two

most likely trade off, and an improvement in one may not always translate into improvement in the other. Because of this, I cannot use any measure related to sequence prediction (e.g. perplexity) to ground the 'good start' quantitatively; instead I use the balanced accuracy (computed as described in chapter 2). The quantitative definition of a 'good start' is relatively straightforward: A 'good start' is a special kind of training experience supplied to an untrained model that results in greater balanced accuracy compared to a model without a 'good start'. But this definition does not provide any insight; I need to develop a qualitative definition to be able to make predictions (which can be tested quantitatively). Previously, I have provided a working qualitative definition as follows: A 'good start' is a special training experience supplied to an untrained model such that the model quickly reaches a location in parameter space that facilitates subsequent acquisition of semantic category knowledge. The goal of this section is to refine this working definition into a theory that is worth testing.

To begin, let's take a closer look at what it means to say that a learning experience 'facilitates *subsequent* acquisition'. One way in which a future learning experience is positively affected by a past learning experience is if the previous learning experience has endowed a learner with knowledge that *generalizes* to a subsequent experience. Generalization is the holy grail of learning, because it enables the system to predict outside of the training data. To apply what one has learned to novel situations, one cannot simply store the training data in memory. Instead, a good learning system is one which learns the *structure underlying the input.* A novel experience is likely to have the same underlying structure, and therefore the learner can use its knowledge about the underlying structure to make predictions in novel situations. With this in mind, one can expand the working definition of a 'good start' as such: A 'good start' guides a model to a location in parameter space which increases the model's ability to generalize to novel

situations (relative to an identical model which did not 'start good'). So far so good; but this definition is still too general. The conditions under which generalization occurs is a hotly debated topic in cognitive science, and numerous conditions and requirements have been identified. I would like to further constrain the definition. To do so, it is worth looking at the input that the SRN actually receives. I have conducted extensive corpus analyses of AO-CHILDES in chapter 4, and found numerous corpus-statistical factors that correlate strongly with age of the target child.

If it is true that training on speech to younger children in AO-CHILDES provides greater generalization of semantic category knowledge, then there must exist some identifiable corpus-statistical change in AO-CHILDES that correlates with age. Because I am talking about generalization of knowledge about semantic categories, it is tempting to think that semantic category structure is simply more clearly defined in speech to younger children compared to speech to older children[13]. In other words, one might wonder whether the semantic category structure itself is systematically changing across AO-CHILDES. Perhaps the distributional cues that define the semantic category structure degrade in quality, such that information about semantic category membership is inversely proportional to the location in AO-CHILDES? These are precisely the kind of questions I investigated in the corpus analyses described in chapter 4. While two out of three analyses suggested that there may be slightly more information about the semantic category structure in partition 1, a third analyses indicated the opposite. Together, these results suggest that there is virtually no difference in the amount of information about semantic category membership between the two partitions. While the *quantity* of the information about semantic category structure does not appear to vary with age of the target child, what about the

[13] When I talk about the semantic category structure of AO-CHILDES, I am referring to the structure defined by the 28 categories used in this work, and described in detail in chapter 2.

quality of the semantic category structure? I don't mean to say that the categories themselves change, but that category-context links in the input might change. For example, it is possible that the set of words that tend to co-occur with members of the category MAMMALS - and therefore distributionally define the category - might systematically change. Specifically, new words could be added, and/or existing words can be dropped. Given the relatively small number of probe word occurrences in each partition (131,236 in partition 1 and 105,670 in partition 2), I find there is little room (and far too much noise) for any systematic differences of this kind. More importantly, claiming that semantic category structure in AO-CHILDES changes qualitatively from one partition to the next implies that there is more than one source of distributional information about semantic category membership available. This violates a core principle of the of language: each entity in the world is referred to by at most one word (with some exceptions). If it is true that there are alternative sources of distributional information that converge on the same semantic category structure, speakers would have to use different words systematically to refer to the same entity. Moreover, they would have to do so as a function of the listener's age. Note, that I am not talking about alternative words for probe words; I am talking about the words that make up the context of probe words (6 words preceding the probe word, see chapter 2). For example, everything else being equal, consistently using *leash* and *bark* in the context of *dog*, would hold the same amount of distributional information about the semantic category of *dog* than consistently using two different words with the same meaning. As said before, there are few words in English which are both frequently used and have the same meaning. While it is possible to use the diminutive alternative form when talking to very young infants (*doggy* vs *dog*), this is a special case restricted to a small set of animate entities. It is almost impossible to come up with other examples of cases where more than one (infant-

compatible) word is used to refer to the same entity. What alternative words exist for *dog*, *plate*, *car*, etc. which do not hamper the ability of the child's vocabulary development? While there are more scientific or formal alternatives (e.g. *canid*, *platter*, *automobile*), these are virtually nonexistent in child-directed speech, and specifically in AO-CHILDES. Another way to approach the possibility that semantic category structure varies qualitatively with age of the target child, is the distribution of topics changes. This is a more plausible way in which semantic category structure could vary across AO-CHILDES, without affecting the quantity of information about the semantic category structure. For example, rather than using two different words to refer to a leash when talking about dogs, *leash* can be a cue to semantic category membership in one partition but another word like *bone* can take on a similar role in the other partition. While plausible, it is very unlikely. Probe words used in this work refer to very common entities, and therefore leave little room for idiosyncratic usage. In other words, the diagnostic context words for *dog*, *cup*, or *car* are unlikely to change as a function of age, because the situations in which they are used are highly stereotypical across the lifespan. If we accept the above arguments, we are left with a somewhat counter-intuitive conclusion: The aspect of the input responsible for interference during SRN training (supposed to explain the differential performance in semantic category learning between models trained in age-order and reverse age-order) is most likely not an incrementally changing semantic category structure.

If not an incrementally changing semantic category structure, what other corpus-statistical factor may explain the greater generalization after having trained on partition 1 first? Is it possible that some difference in the *overall sequential structure unrelated to probe words* can give rise to a 'good start'? In other words, does some structural property of AO-CHILDES (e.g. syntactic structure), that does not alter the semantic category structure, influence the SRN's

representation of the semantic category structure? This question is tested empirically in chapter 7; however, there are several a priori reasons why this should be the case. For example, it is worth noting that there are only 532 probes, less than 20% of the size of the vocabulary. Any systematic change in the remaining 4,096 - 532 = 3,564 non-probe words in the vocabulary has a large impact on the SRN's evolving representational space, which may indirectly affect the representations of probe words. The mere fact that non-probe words represent an overwhelming majority in the vocabulary (and in the input) is only one reason that I am interested in further investigating the effect of incrementally changing syntactic complexity on the semantic category structure of probe words. The primary reason is the pattern of results obtained in the corpus analyses. I found that the syntactic complexity of AO-CHILDES is gradually increasing, while information related to the semantic category structure remains more or less constant. Taking this finding seriously requires us to consider that the SRN's acquisition of semantic knowledge may be affected by the incremental change in the syntactic complexity of the input. But, how can the reduced complexity in the syntactic structure of partition 1 improve generalization of semantic category knowledge? After all, the goal is to define exactly what it means to 'start good' and I already said that it must have to do with a better ability to generalize to novel examples of the semantic category structure. This question is the subject of the remainder of this chapter. Before moving on, let's update our working definition: A 'good start' is provided by input with less complex syntactic structure.

I am unaware of previous studies addressing the interaction between semantic and syntactic structure in the SRN, and therefore cannot refer the reader to existing evidence that such an interaction even exists. But it is relatively straightforward to explain why the answer should be *yes*. Simply put, the SRN does not distinguish a priori between these two types of

structures. Both syntactic and semantic phenomena in the input are fed to the model in identical fashion, and stored in the same set of weights. In fact, a single weight change is never exclusively due to syntactic or semantic structure in the input; each weight update incorporates information present in 64 windows, each 7 words long (a mini-batch). Even a single three word utterance like *the dog barks* contains both semantic (*dog* co-occurs with *bark* more often chance) and syntactic structure (verb follows noun). In some linguistic theories, syntactic and semantic structure are determined by two independent systems, and a model of acquisition provided with knowledge about these two systems would, in principle be able to encode syntactic and semantic dependencies in the data separately, without the possibility of cross-talk. But the SRN does nto distinguish between semantic and syntactic structure, and therefore cannot encode each separately. Because no such separation exists, any change to the representational space could modify encodings of both structures.

As an aside: Whether the SRN is actually capable of encoding syntactic structure given a finite sample of natural language is still debated, and the answer in large part depends on one's theory of syntax. That said, I use the phrase "encoding syntactic structure" as a shorthand for "encoding relationships in the data which approximate the underlying syntactic structure of the input". The goal of this work is not to address whether the SRN can faithfully encode the syntactic structure underlying English child-directed speech, but to investigate the learning dynamics of the SRN when learning from input with multiple (possibly independent) structures.

Grammatical categories are acquired first

I now return to the question of how syntactic structure can influence the semantic category structure that the SRN acquires. To glimpse ahead, I will show that the SRN must

acquire semantic categories by carving them out of an existing syntactic category. This carving is where representation of syntactic and semantic category structure interact in the SRN. To develop this argument, I must first explain why I think that the SRN, when trained to predict natural language sequences, first acquires syntactic categories (e.g. NOUN, VERB) before semantic categories (e.g. MAMMAL, VEHICLE, TOY). I will borrow extensively from Saxe, McClelland & Ganguli (2019) who have provided formal evidence that a deep linear neural network exhibits progressive differentiation, a process characterized by learning gradually finer-grained aspects of the input. Progressive differentiation is best illustrated when the input to a learning system consists of a set of items (or concepts) each associated with a set of semantic properties (e.g. the item *penguin* has the properties *has_legs* and *can_swim*). In the real world, such data is often hierarchically structured, and therefore the properties that *penguin* and *canary* have in common do not needed to be stored separately; rather, information about both *penguin* and *canary* can be stored at a superordinate level (e.g. the category BIRDS). Further, properties shared between members of both BIRDS and, say MAMMALS, can be stored at yet a higher level (e.g. ANIMALS) encapsulating both subordinate categories. Further grouping of categories by some measure of similarity (e.g. the number of features two categories have in common), would result in a hierarchy, where the highest level branch might divide ANIMATE from INANIMATE, for example. When such hierarchically structured data is provided to a deep linear neural network, Saxe, McClelland & Ganguli have shown that the network first encodes the top-most distinction of the hierarchy (ANIMATES vs. INANIMATES), followed by subordinate category distinctions (e.g. ANIMALS vs. PLANTS) and only then learns to differentiate between individual items (e.g. *penguin* vs. *bird*). Given infinite resources and time, a network will eventually memorize the input data. The authors have suggested that this learning dynamic need

not be specific to linear networks, and is similar to that observed in nonlinear networks studied by Rumelhart and Todd (1993), and Rogers & McClelland (2008). I will assume the same dynamics are applicable to the SRN (which, when unfolded in time, is identical with a deep feed-forward neural network).

But what does this mean for input with sequential structure, like natural language? The theory developed by Saxe, McClelland & Ganguli (2019) predicts that the SRN's predictions are initially guided by the top-level category of the word in the input. Given a probe word in the input, this means that the SRN is primarily (and possibly, exclusively) using the fact that the probe word is a noun to constrain next-word prediction, during the early stages of training. Knowing that the word is a noun vastly reduces the number of possible answers, because English nouns are consistently followed by verbs. The theory also predicts that semantic distinctions between individual nouns are learned later, because these distinctions occur less frequently and therefore explain a smaller proportion of the variance in the input (as determined by singular-value decomposition).

Because learning in a neural network follows a quasi-stage-like trajectory, in which the dimensions of the input are acquired in order of the variance they explain (as determined by singular-value decomposition, or SVD), it is possible to predict which dimensions the SRN is likely to learn first. Of primary interest is determining whether the earliest acquired dimensions account primarily for syntactic or semantic variance. This can be done by constructing a matrix where columns represent all unique 7-word windows the SRN is fed as input, and where rows represent the word following the 7-word window in AO-CHILDES. The results is a matrix with 4096 rows and 4,451,458 columns, and I will refer to it as the term-by-window co-occurrence matrix. Each element in the matrix is assigned the frequency with which the 7-word window

represented by the column precedes the word represented by the row. Because of the large number of columns, I restricted the analysis to the three singular dimensions with the largest singular value. The results are overwhelmingly in favor of the idea that syntactic categories hold greater explanatory power than semantic categories: Words loading highest on the largest singular dimension are *read, take, put, help, open, hold, make, get,* which are all verbs. There does not appear to be any organization of verbs that reflects their semantics. The same words load highest on the second largest singular dimension, but this dimension clearly distinguishes verbs from interjections like *uhuh, well, okay, mhm, yeah, yes, no*. Lastly, words loading highest on the third largest singular dimension are *maybe, where, I, let, we, but, why,* which are all frequently found at utterance-initial positions, following punctuation. These words are markers of utterance boundaries, and therefore serve little, if any semantic function. These results align well with the PCA analysis of the SRN hidden states described in chapter 2. There, it was found that the first principal component encodes whether a word can occur in isolation, and that the second principal components encodes roughly the distinction between NOUN vs. VERB. Both dimensions appear to encode purely syntactic variance in the data, and explain roughly 30% of the variance in the hidden states.

Given that syntactic categories make up the majority of the most prominent singular dimensions, one of the earliest categories that the SRN must learn is the category NOUN. This category reflects the SRN's knowledge about words that occur in noun-like contexts. I will keep referring to this category using capital letters to indicate that this category may or may not reflect a linguist's definition of the syntactic category. Because the probe words used in this work are all nouns, all semantic categories acquired by the SRN must therefore originate in the category NOUN. The success of the semantic differentiation of the category NOUN - and consequently,

subsequent performance on semantic categorization - will largely depend on how the category NOUN was initially encoded. This predicts that the semantic categorization performance should be greater during very early stages of training in age-order, when the model is acquiring knowledge about the syntactic structure underlying the input (e.g distinguishing nouns from verbs). Indeed, this is the case: a semantic categorization performance gap is apparent as early as one tenth of the input has been trained on (see chapter 3).

At this point it might be worth explaining in more detail why semantic categories whose members are nouns, must originate in the SRN's NOUN category. Let me illustrate with the following example:

(a) the book is

(b) the man is

(c) the dog __

In all three utterances a, b, and c, the determiner *the* is followed by a noun, and in both cases (a) and (b), the noun is followed by the verb *is*. Say that the SRN has previously seen examples a and (b), but has never before encountered (c). In this case, the SRN will predict *is* in alignment with its previous experience. Why? Because, although *dog* has never before been seen, the activation of *the* at the hidden layer contributes a pattern of activations at the hidden layer at the next time step that suggest that next word is likely going to be a member of the previously acquired category NOUN. The knowledge that *is* tends to be followed by a NOUN is encoded in the recurrent weights. The activation of *is*, in combination with the recurrent weights, results in a pattern of hidden activations at the next time step that resembles that of a NOUN. Given the NOUN-like activation at the hidden layer, the unit coding for *is* will be strongly activated at the output layer. If the correct next word is in fact *is*, then the weight updating procedure will modify

the representation of *dog* in the direction of the NOUN template. If, however, the correct next word is bark, which does not systematically follow all members of the category NOUN, then the representation of *dog* is updated to reflect it is part of the NOUN category, but with a slight departure from the NOUN template. Put differently, *dog*, by virtue of occurring in a NOUN-like context will be assigned a NOUN-like representation at the hidden layer, regardless of the identity of the correct next word. If the correct next word is consistent with a NOUN interpretation, then the NOUN status of *dog* is reinforced. If, however, the correct next word is unique to dog (or its semantic category), the representation assigned to *dog* will be a departure from the NOUN template. In either case, the weights assigned to the input unit coding for *dog* (initially random) are made to resemble those coding members of the category NOUN. The only possibility in which the representation of *dog* is not made more NOUN-like is if *dog* did not occur in a NOUN-like context. This is an unlikely possibility in AO-CHILDES.

A theory of the age-order effect

With several pieces of the puzzle in place, I will outline a theory with the purpose of explaining the mechanism underlying the age-order effect described in chapter 3. The theory is composed of several arguments, some of which are supported by evidence in this work, and some of which are supported by studies conducted by other researchers. After a brief overview of the arguments below, I will discuss each in detail in the following section. Argument 1 is skipped because it has been discussed in detail in chapter 4.

1. The surface structure of partition 2 of AO-CHILDES is more complex compared to partition 1. Specifically, there is a greater variety of constructions, which is associated

with longer utterances, and greater density of function words like conjunctions and prepositions which connect clauses into complex utterances (see chapter 4).

2. A consequence of the increased number of constructions is that the sum of the column variances of the term-by-window co-occurrence matrix computed on partition 2 is smaller compared to partition 1.
3. A decrease in the sum of the column variances of the term-by-window co-occurrence matrix, while keeping constant the sum of all elements in the matrix (as must be the case given that both partitions are of equal length), is accompanied by an increase in the sum of the singular values.
4. Because the amount of training time required to acquire a singular dimension in a deep (linear) neural network is inversely proportional to the magnitude of the associated singular value (Saxe, McClelland & Ganguli, 2019), SRN training on partition 2 is marked by longer periods of overlap between acquisition of different singular dimensions. This may blur the SRN's distinction between semantic and syntactic categories during training on partition 2.
5. The representational blurring during training on partition 2 is reduced by training in age-order because the SRN has established more distinct and stable semantic and syntactic categories during training on partition 1. The benefit of age-ordered training, is that it reduces the problem of overlapping periods of semantic and syntactic differentiation.

Argument 2

I argued that the greater number of unique constructions in partition 2 is accompanied by a decrease in the column variances of the partition 2 term-by-window co-occurrence matrix

relative to partition 1. Let me illustrate this using a simple example. Let TW_1 and TW_2 define two term-by-window co-occurrence matrices of size *n* by *d* where *n* is the number of words and *d* is the number of windows in a hypothetical corpus.

$$TW_1 = \begin{bmatrix} 10 & 0 \\ 9 & 1 \\ 0 & 10 \\ 1 & 9 \end{bmatrix} \quad TW_1 = \begin{bmatrix} 7 & 3 \\ 6 & 4 \\ 3 & 7 \\ 4 & 6 \end{bmatrix}$$

Suppose the two term-by-window co-occurrence matrices were constructed from two partitions of the same hypothetical corpus, which I will call p_1 and p_2. Both partitions are of equal length, and this can be verified by comparing the sum of TW_1 with the sum of TW_2. In both cases, the sum is 20. This means there are twenty words, or 20 windows in each partition. While the number of tokens is identical, the number of types (of constructions) may be different - and in fact, this is the important difference here. Put differently, the number of unique constructions in p_1 is smaller compared to p_2. This can be verified by comparing the sum of nonzero elements in TW_1 and TW_2 (6 vs. 8, respectively). The greater the number of nonzero elements, the greater is the number of term-window combinations. Consequently, a larger number of zero values indicates that some term-window combinations do not occur in the partition. But I have not yet shown that the sum of the column variances are larger for p_1 compared to p_2. This can be verified by computing the sample variance of each column vector in TW_1 and TW_2 and comparing their sum. The result is in agreement with my argument: Because the column means are 5.0 (column 1) and 5.0 (column 2) for both matrices, the column variances are:

$$VAR(TW_{1,column1}) = \frac{(10-5)^2 + (9-5)^2 + (0-5)^2 + (1-5)^2}{4} = 20.5$$
$$VAR(TW_{1,column2}) = \frac{(0-5)^2 + (1-5)^2 + (10-5)^2 + (9-5)^2}{4} = 20.5$$
$$VAR(TW_{1,column1}) = \frac{(7-5)^2 + (6-5)^2 + (3-5)^2 + (4-5)^2}{4} = 2.5$$
$$VAR(TW_{1,column2}) = \frac{(3-5)^2 + (4-5)^2 + (7-5)^2 + (6-5)^2}{4} = 2.5$$

The sum of the first two terms is larger than the sum of the last two. Keeping the sum identical but varying the sum of the column variances, has important implications for the singular-value decomposition of the term-by-window co-occurrence matrix.

The value of this demonstration may not be obvious. The two matrices TW_1 and TW_2 were chosen to exemplify the term-by-window co-occurrence matrices of partition 1 and 2 of AO-CHILDES, respectively. In p_1, the number of unique constructions is smaller relative to p_2, and this same pattern is true for AO-CHILDES. This was achieved by setting more cells in TW_1 to zero, which indicates that those term-window combinations do not occur. To maintain identical sums between TW_1 and TW_2, however, the remaining nonzero values in TW_1 must not be the same as the corresponding values in TW_2. To satisfy this constraint, I created TW_1, and showed that the sum of its column variances must be greater compared to TW_2. By extension, the same logic may be applied to the term-by-window matrices for AO-CHILDES partitions.

Argument 3

Next I want to provide a technical argument for why the sum of the singular values of two term-by-window co-occurrence matrices, whose sums are kept constant but whose column variances differ, cannot be identical. Specifically, I argued that the sum of the singular values associated with the matrix whose sum of column variances is larger compared to the matrix

whose sum of column variances is smaller. To understand why this is true, let's return to the example of the hypothetical corpus, and the two term-by-window co-occurrence matrices, TW_1 and TW_2. There are two 'distributional' categories, A and B, in the corpus. In both TW_1 and TW_2, the first two rows represent words that are members of category A,while the second two rows represent words that are members of category B. In both partitions, category A words co-occur more frequently with the window represented by the first column, and category B words co-occur more frequently with the window represented by the second column. This is the 'distributional' definition of the two categories. Singular-value decomposition is a tool highlights the most prominent dimensions underlying high-dimensional data. In this case, the example data only has 2 dimensions, for clarity. Nonetheless, it will be informative to walk through the process of how SVD applies in the 2-dimensional case. The task of SVD, here, is to recover a dimension that distinguishes between the two categories A and B. Because we know that there are only two categories a priori, we only need one singular dimension to explain the category structure of the corpus, and discard the second singular dimension. A second singular dimension, however, will be detected by SVD, and will encode information about the difference in the frequency of the term-by-window co-occurrences of the two words within each category. I purposefully varied the pattern of co-occurrence frequencies slightly between words that are in the same category to be able to observe and discuss any difference between the singular values associated with this second singular dimension. If the two words in each category had the same co-occurrence pattern (their rows in TW_1 and TW_2 would be identical), then no second singular dimension would be detected. We will see that the crucial difference, in the result of SVD, is reflected in this second singular dimension.

One way to understand how singular values are obtained is by visually examining the data. In the left two panels of figure 6.1, the points represent the row vectors in TW_1 and, in the right panels, the points represent the row vectors in TW_2. Each word is a point in a 2-dimensional distributional space. Each dimension represents a word's frequency of occurrence with window 1 (x-axis) and window 2 (y-axis). The first singular value, s1, is computed by first projecting each point onto the best-fit line which must cross the origin (dashed grey line in top panels). There are four projections in total, one for each word, and they are shown in red. (The projections are slightly offset to show there are four in each panel.)

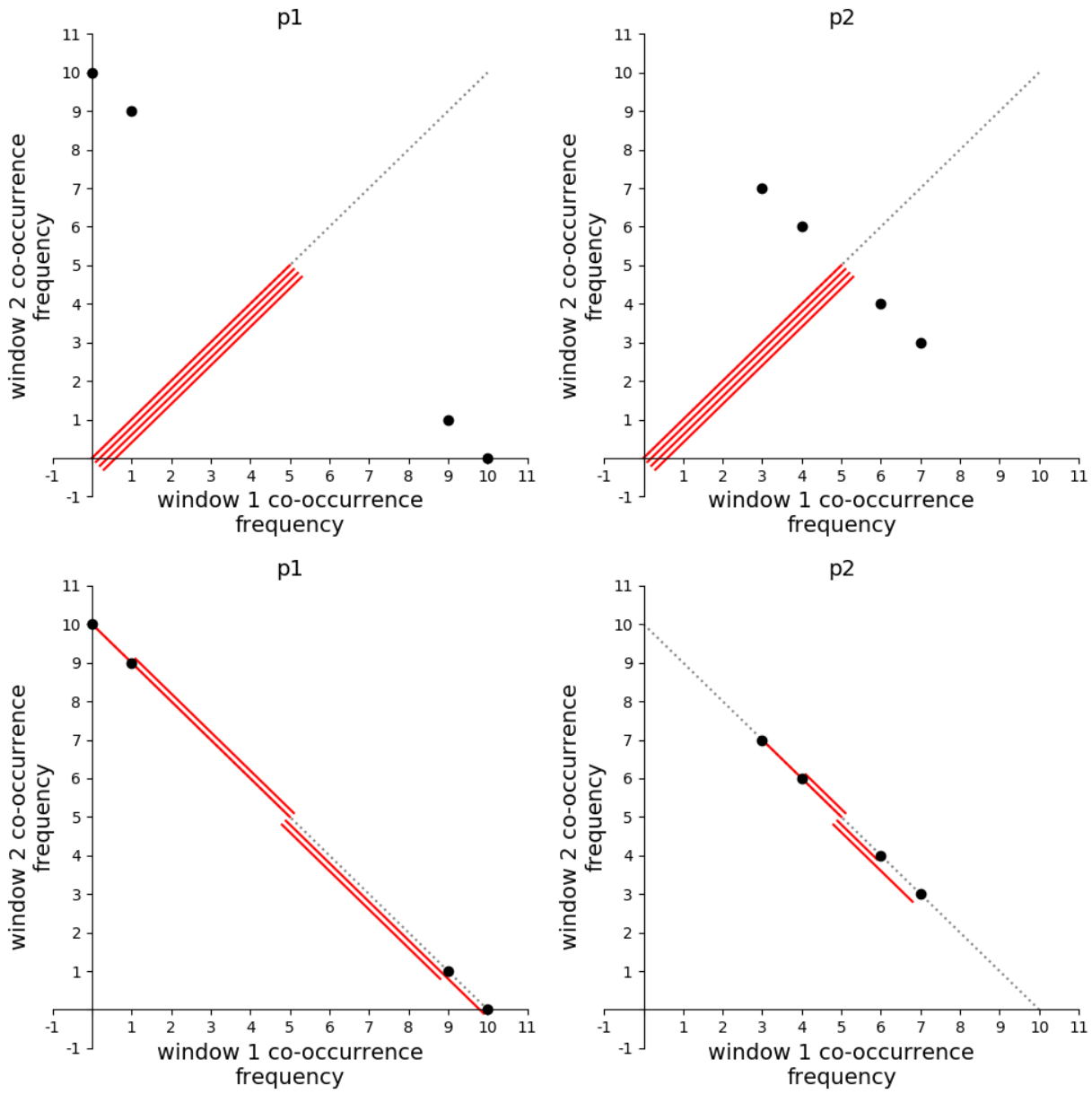

Figure 6.1
Plot of term-by-window co-occurrence data for p_1 (left panels) and p_2 (right panels) of hypothetical corpus with 20 sequences and a vocabulary size of 4. Each dot represents a word in distributional space. A word is defined by its corresponding row in the term-by-window co-occurrence matrix. In this case there are 2 columns, meaning a word can vary along 2 dimensions: How often it occurs with context 1 and context 2. The top row illustrates computation of singular value 1, and the bottom row illustrates computation of singular value 2. Red lines are projections of dots onto a best-fit line indicated by the dashed grey line. The best-fit line in the top panels must cross the origin, and the best-fit line in the bottom panels must be orthogonal to the best-fit-line in the top panel.

Next, a singular vector is obtained, v_1, which is the best-fit line scaled to unit-length. Thus,

$$v_1 = \begin{bmatrix} 0.70710678 \\ 0.70710678 \end{bmatrix}$$

The final step is to actually compute s_1. This can be achieved by calculating the sum of all the squared projections of each point onto s_1. Using the tools of linear algebra, this can be done by computing the L_2 norm of the dot product of the data matrix with v_1, as such:

$$s_1(TW_1) = \|TW_1 v_1\|_2 = 14.14$$

$$s_1(TW_2) = \|TW_2 v_1\|_2 = 14.14$$

The result is:

$$s_1(TW_1) = \left\| \begin{bmatrix} 10 & 0 \\ 9 & 1 \\ 0 & 10 \\ 1 & 9 \end{bmatrix} \begin{bmatrix} 0.70710678 \\ 0.70710678 \end{bmatrix} \right\|_2 = 14.14$$

$$s_1(TW_2) = \left\| \begin{bmatrix} 7 & 3 \\ 6 & 4 \\ 3 & 7 \\ 4 & 6 \end{bmatrix} \begin{bmatrix} 0.70710678 \\ 0.70710678 \end{bmatrix} \right\|_2 = 14.14$$

The first two singular values are identical. This is expected, because their associated singular dimension does not encode the category distinction A vs. B, but the magnitude of the vectors that define each dot (or, their distance from the origin). All row vectors in the data matrix have equal magnitude, therefore this singular dimension is not informative about the distributional structure of the corpus. However, we will see that the second singular value, v_2,

which represents the category structure (A vs. B), and, importantly, is larger for the data matrix computed on p1 compared to p_2. The formulas are the same, except that the best-fit line does not cross the origin, and must instead be orthogonal to the previous best-fit line (in top panels). To obtain s_2, calculate:

$$s_1(TW_1) = \left\| \begin{bmatrix} 10 & 0 \\ 9 & 1 \\ 0 & 10 \\ 1 & 9 \end{bmatrix} \begin{bmatrix} +0.70710678 \\ -0.70710678 \end{bmatrix} \right\|_2 = 12.81$$

$$s_1(TW_2) = \left\| \begin{bmatrix} 7 & 3 \\ 6 & 4 \\ 3 & 7 \\ 4 & 6 \end{bmatrix} \begin{bmatrix} +0.70710678 \\ -0.70710678 \end{bmatrix} \right\|_2 = 4.47$$

This result can be verified by visually inspecting the length of the projections (shown in red) in the lower two panels. The projections of dots (onto the best-fit line indicated by the grey line) representing the distributional pattern of words in p_1 (left) are larger than the projections of dots representing the distributional pattern of words in p_2 (right). More importantly, this result is a direct consequence of the larger sum of the column variances of the term-by-window co-occurrence matrix associated with p_1 compared to p_2. The only way in which the projections in the left panel can be larger than those in the right panel, while keeping the sum of the two data matrices constant, is by increasing the column variances. I have shown in the previous section that one way to increase the sum of the column variances is to make the term-by-window co-occurrence matrix more sparse; this would reflect a corpus partition which has fewer term-context combinations.

Argument 4

Next, comes the most crucial aspect of my theory. The above arguments may be viewed as arcane mathematical phenomena, but how do they relate to semantic category learning in the SRN? Remember that Saxe, McClelland & Ganguli have provided a formal description of the learning dynamics of deep (linear) networks, which says that a network acquires singular dimensions of the input in distinct (possibly overlapping stages). According to their formal description, acquisition duration is a highly nonlinear function of the singular value associated with the singular dimension to be acquired by the network. The larger the singular value, the steeper (and earlier) is the acquisition of the singular dimension by the network. The consequence of a larger singular value is not only that its associated singular dimensions is acquired faster (and earlier), but that there is a smaller chance that the acquisition period will overlap with acquisition periods of other singular dimensions. For example, if a syntactic distinction in the input is associated with a large singular value, then the period of acquisition by the SRN is at less risk of overlapping with other periods, say in which semantic differentiation occurs. This would improve semantic categorization performance because the representation of semantic categories will be less influenced by changes to the representation of syntactic structure.

The only unsupported claim is that the learning dynamics of the SRN actually are in agreement with those described by Saxe, McClelland & Ganguli, who studied only deep linear networks. The SRN I use squashes activations at the hidden layer through a nonlinear sigmoid function, making it a deep nonlinear network. In chapter 7, I explicitly test this assumption, and conclude that the SRN's learning dynamics indeed match those described by Saxe, McClelland & Ganguli.

Argument 5

The argument that age-ordered training reduces the problem of overlapping periods of differentiation follows directly from arguments 1-4. Because SRN training on partition 1 of AO-CHILDES establishes more distinct category representations, it should be less affected by the blurring of category distinctions that result when training on partition 2 first. Further training on partition 2 involves refinement of already existing semantic categories, rather than acquisition of new syntactic distinctions. Syntactic dependencies encountered in partition 2 have little effect on the representational space of the SRN, because syntactic categories have by then been firmly established. It is possible that some syntactic dependencies only exist in partition 2, which may influence the course of semantic differentiation during age-ordered training on partition 2, but their influence would be relatively small. Presumably, by the time training on partition 1 has been completed, the SRN has already mastered the most frequent syntactic distinctions in AO-CHILDES. During partition 2, the SRN can further refine semantic categories without competing influence by other non-semantic distinctions.

An Example: Trajectories through Representational Space

To arrive at a more complete understanding of the theory described above, I turn to figure 6.2 which shows a hypothetical trajectory of four-word representations through the SRN's hidden (or, input) state space over the course of training. Panel A is supposed to demonstrate the semantic differentiation that should (According to my theory) occur during training on partition 1 of AO-CHILDES, while panel B is supposed to demonstrate the same process during training on partition 2. Before describing the figure, a quick note: Both figures were drawn without having inspected any actual state space trajectories (which might be obtained by using multi-

dimensional scaling). As such, these drawings represent a relatively unbiased view of how competition between semantic and syntactic category distinctions might look like. We must also keep in mind the figure was drawn for pedagogical purpose only. It might be useful to compare them to the actual state space trajectories but it is not obvious whether exactly the same pattern should be observed. Competition between semantic and syntactic categories can manifest in numerous ways, and my theory cannot be used to predict how exactly competition will manifest and for which words.

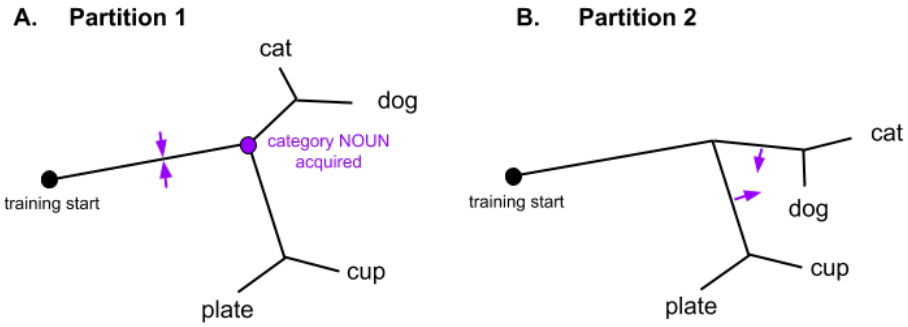

Figure 6.2
Hypothetical trajectories of 4 words through the SRN's hidden state space during training on partition 1 (left panel, A) and partition 2 (right panel, B). Purple arrows indicate the force on representational trajectories exerted by acquisition of the category NOUN. Crucially, this force influences semantic differentiation during training on partition 2 (purple arrows pull representations of cat and dog, and plate and cup closer together) but does not influence semantic differentiation during training on partition 1.

Now, in both panels, the start of training is indicated by the black dot on the left of each panel. The black lines indicate movement of the representations for *cat*, *dog*, *cup*, and *plate*. Initially movement of all four words is identical, because early training is characterized by acquisition of syntactic category distinctions. In this case, all four words are nouns, and the movement of the representation of the 4 words through hidden state space is initially influenced

exclusively by their membership in the category. The purple dot in panel A indicates the time point at which acquisition of the category NOUN is completed and differentiation of the semantic categories ANIMAL vs KITCHEN ITEMS begins. Notice there is no purple dot in panel B. There, acquisition of category NOUN continues, and importantly, overlaps with semantic differentiation of ANIMALS vs. KITCHEN ITEMS. During semantic differentiation, cat and dog are moving farther away from plate and cup. This process happens in isolation in partition 1 (panel A), but happens in parallel with the continuing acquisition of the category NOUN in partition 2 (panel B). The purple arrows indicate the force exerted by acquisition of the category NOUN, which is pushing all 4 representations closer together. This is not a problem during training on partition 1 where this force ceases to exist at the time point indicated by the purple dot; however, this is a problem during training on partition 2 where this force can influence semantic differentiation which is occurring concurrently. Here, the force exerted by acquisition of the category NOUN is pulling ANIMALS and KITCHEN ITEMS closer together (as indicated by the purple arrows in panel B). The result, at the end of training is that the representations of cat and dog are more similar to plate and cup in partition 2 compared to partition 1. Concurrent acquisition of the syntactic category NOUN has partly reversed the beneficial effect of semantic differentiation, resulting in less distinct representations for the two semantic categories.

I have made it sound as if there was a clear point, during training, at which the category NOUN has been acquired. But the situation is clearly more complicated. To be precise, acquisition of a singular dimension is never completed. Acquisition of a singular dimension, as demonstrated by Saxe, McClelland & Ganguli (2019) follows a sigmoidal trajectory, which does not reach its asymptotic value in finite time. Nonetheless, because the trajectory is sigmoidal,

there is a recognizable transition period in which performance rapidly increases to near the asymptotic value. More importantly, I have evidence that demonstrates clearly that there is a point, during training, at which syntactic differentiation is no longer detectable. Instead of evaluating the balanced accuracy for semantic categories, I have evaluated the balanced accuracy for a number of syntactic categories (NOUN, VERB, ADJECTIVE, DETERMINER, ADVERB, PREPOSITION, CONJUNCTION) and tracked performance across training. At approximately 1/8th of the total number of training steps until completion, balanced accuracy for syntactic categories has reached its asymptotic value. Put differently, differentiation of syntactic categories is (near-to) complete shortly after training has begun. In contrast, semantic differentiation takes place throughout the entire duration of training; balanced accuracy for semantic categories only reaches its asymptotic value at the very end of training.

Support for an effect of syntactic complexity on semantic category learning

An important prediction of the theory is that syntactic complexity is the key driver of the incremental structural change in AO-CHILDES. As such, an ordering of AO-CHILDES, not by age, but by some measure of surface structure complexity, should recapitulate an order effect that closely resembles the age-order effect. To do so, I partitioned AO-CHILDES into 256 chunks and ordered them by entropy of the discrete frequency distribution of words in a chunk. I trained two groups of SRNs: one on partitions ordered by increasing entropy, and another on partitions ordered by decreasing entropy. If it is true that complexity of surface structure is the driver of the age-order effect, then a semantic categorization performance gap between the two groups, that is at least as large as that observed when ordering partitions by age, should be observed. Results showed a strong performance gap, that closely resembles the age-order effect, but with an even

larger gap (x2) at the end of training. This is strong evidence in favor of the idea that incremental change in complexity of surface structure is the driver of the age-order effect.

What evidence exists that a distributional semantic model, such as the SRN, is affected by syntactic structure in the input? Baroni & Lenci (2011) found that inclusion of function words in a simple matrix-factorization based distributional model resulted in greater cosine similarity of random (semantically unrelated) noun word-pairs. On the other hand, exclusion of function words greatly reduced clustering of nouns. This is not surprising, because function words like determiners and articles consistently co-occur with nouns, and therefore provide evidence to the model that nouns should form a category. While this kind of clustering is useful if the goal is to learn the syntactic structure of the input, when semantic categories are desire, function words are instead providing conflicting distributional information.

More support for the notion that syntactic category influences category membership comes from Huebner & Willits (2018). Briefly, the SRN exhibited a taxonomic bias, meaning that neighbors in the SRN's hidden state space are words referring to entities that tend to be of the same kind rather than entities that tend to occur in the same event. This result need not be true of all distributional semantic models, and in fact, Huebner & Willits (2018) showed that Word2Vec, which is only minimally constrained by word order does not exhibit a taxonomic bias. Rather, neighbors in Word2Vec hidden state space tend to be thematically (e.g. dog, bark, leash), as opposed to taxonomically related. This may make Word2Vec relatively immune to order effects related to the syntactic complexity of the input on which it is trained.

The size of the context windows that the SRN is trained with, limits the complexity of the surface structure in the eyes of the SRN. Training on larger context windows means the SRN sees a greater number of unique windows. As long as the context size is set larger than the

largest utterance length, the SRN is guaranteed to see the full syntactic complexity of the input (excluding syntactic dependencies that cross utterance boundaries). The context size used in this work is 7 which is close to the average utterance length in AO-CHILDES. My theory predicts that training on smaller context windows should reduce the age-order effect, because smaller context windows reduce the impact of increasing structural complexity in AO-CHILDES on the SRN. Indeed, training two groups of SRNs as was done in chapter 3, but with context window sizes ranging from 2 to 7, I found that the age-order effect is only recognizable for models trained with context window sizes of at least 4. Semantic categorization performance was indistinguishable at the end of training between groups of models trained on context window sizes smaller than 4. Performance differences were detected at earlier stages during training, but they did not persist.

Another crucial component of my theory is that semantic and syntactic structure can compete for representational space. This occurs during overlapping periods of semantic and syntactic differentiation. It is important to keep in mind however that this is not the only scenario in which competition can occur; in fact, parallel differentiation of multiple semantic distinctions may influence the representational space in unpredictable ways. Similarly, parallel differentiation of multiple syntactic distinctions can have the same effect. In either scenario, my theory predicts that a smaller SRN, with decreased representational capacity, should be more prone to competition. One way to reduce the size of the SRN is to reduce the number of units at the hidden layer. The effect of training smaller SRNs on semantic categorization performance should be negative, but more importantly, the age-order effect should be greater than that observed for a larger SRN. Toward this end, I trained two groups of SRNs with 128 or 512 hidden layer units on either age-ordered or reverse age-ordered AO-CHILDES, and compared end of training

performance. Indeed, while the performance of the smaller SRNs was overall reduced compared to the larger SRNs, the performance gap between the two training conditions was larger for the smaller compared to the larger SRNs. But not so fast. Training intermediate-size SRNs (256 hidden units) did not result in a greater age-order effect; in fact, the age-order effect was reduced. Clearly, the effect of representational capacity on performance is not as clear cut as I had expected. The results are more consistent, however, when increasing the vocabulary size, which results in a concomitant increase in the number of input and output units. The age-order effect was observed for a vocabulary of size 4,096. Using a vocabulary twice as large (8,192) and four times as large (16,384), I found that semantic categorization performance drops consistently with vocabulary size (from 74.5 to 73.2 and 73.0 for vocabulary sizes 4,096, 8,192, and 16,384, respectively). A larger vocabulary size increases structural complexity of the input by increasing the number of unique constructions possible. Rather than needing to predict the OU-OF-VOCABULARY symbol at the output layer, the SRN is burdened by having to predict the exact identity of the word. Reduced semantic categorization under such circumstances is to be expected under a theory that casts too much structural complexity as a counter-force to semantic differentiation.

More evidence in favor of syntactic effects on semantic categories comes from randomization of word order within context windows that the SRN sees. In such a scenario, the SRN's task is to predict the next word given a shuffled set of words which precede the target word. Shuffling of context windows eliminates the ability of the SRN to use word-order in predicting an upcoming word. Results of training SRNs with shuffled context windows show that semantic categorization performance is improved during training on partition 2 (in both age-order and reverse age-order conditions) and left unchanged during training on partition 1. This

demonstrates that partition 2 induces a stronger reliance on word-order cues compared to partition 1 in order to predict an upcoming word. When this information is removed, the SRN is no longer bound to word-order and is presumably more flexible to acquire semantic dependencies which it might not have paid attention to otherwise during training on partition 2.

Support for differences in progressive differentiation

Because my theory relies on the singular values obtained via singular-value decomposition of the term-by-window co-occurrence matrix of partition 1 to be greater than those of partition 2, the first place to start testing my theory is to actually compute the singular values. The results are shown in figure 6.3, in which singular values are plotted in order of decreasing magnitude starting with the largest and ending in the 64th largest singular value. The results are as predicted: Singular values are larger for partition 1 compared to partition 2. To be precise, I did not predict that *all* singular values would be larger, but that *on average* they would be larger (or, that their sum would be larger) for partition 1. As an aside, the singular values for partition 1 continue to be larger than those of partition 2 well past the largest 64, the cutoff chosen in the figure below. However, the difference shrinks as singular values decrease. I chose to show the first 64 singular values only to highlight the striking difference for the most influential singular dimensions. In sum, the results provide strong support that the ideas developed in this chapter are on the right track.

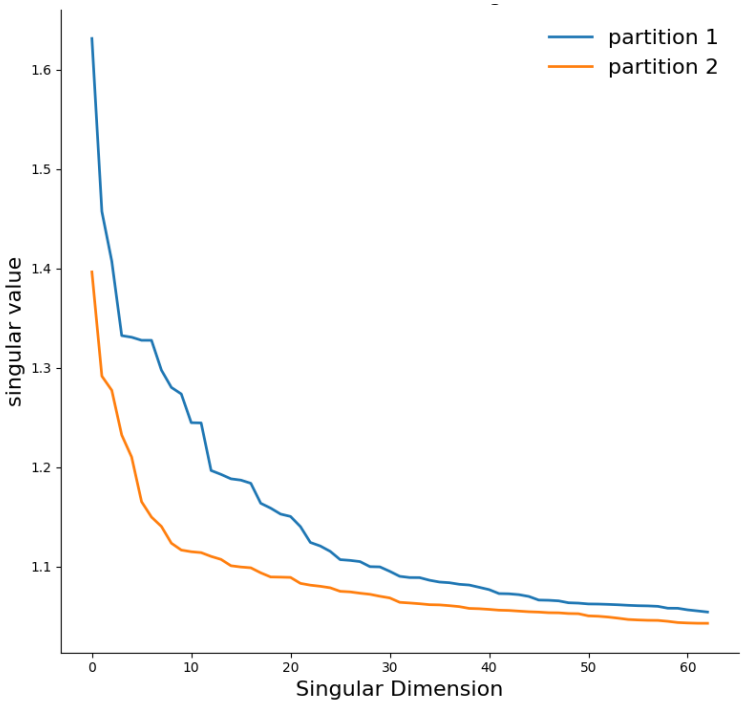

Figure 6.3
Comparison of singular values associated with the first 64 singular dimension obtained via Singular-value decomposition of the term-by-window co-occurrence matrix of partition 1 (blue) and partition 2 (orange)

Next, let's focus specifically on nouns. I have said previously that differentiation of nouns should take longer during reverse age-ordered training (during partition 2) than age-ordered training (partition 1). Before looking at a measure of noun differentiation in the model, let's first inspect the distributional similarity of nouns in the input. Why? The distributional similarity nouns in the input is indicative of the ease with which a distributional learning system would be able to acquire the category. After all, if nouns are distributionally more similar in the input, then they should be more similar in a model's representational space after completion of training. Figure 6.4 shows how the distributional similarity of nouns changes from partition 1 of AO-CHILDES to partition 2. To compute this measure, I first obtained all rows in the term-by-window co-occurrence matrix, computed on a single partition, corresponding to nouns. Next, co-

occurrence representations of nouns were reduced in dimensionality to 512, which corresponds to the number of hidden units used in the majority of simulations used in this work. Lastly, I computed pairwise cosine similarities between these reduced-dimensionality noun representations, and plotted their average (one for each partition). My theory requires that the noun category (and possibly other syntactic categories) is acquired more quickly during training on partition 1 compared to partition 1 (which reduces subsequent competition with semantic differentiation), and therefore the distributional similarity among members of the category should be higher in partition 1 compared to partition 2. The results are well in alignment with this prediction: The cosine similarity is approximately two times as large in partition 1 compared to partition 2 (0.107 vs. 0.062). What exactly might this mean? At this point it would be useful to think again about differentiation as movement of word representations in the SRN's hidden state space: Learning that two words belong to the same category makes them move towards each other; learning that two words belong to different categories makes them move farther apart. During semantic differentiation, both forces have an integral role to play. On the one hand, words which are distinguished on some semantic dimension (e.g. MAMMALS vs. BIRDS) must move farther apart. On the other hand, they must maintain a healthy proximity to each other to maintain their membership in the same superordinate category (if such a category exists, e.g. NOUN). If the latter counter-force is too strong, it may slow the progress of semantic differentiation. In fact, if superordinate category membership is not of crucial importance, it would be best to eliminate the counter-force altogether. In the case presented here, in which the desired outcome is knowledge of semantic, and not syntactic category structure, it would be clearly beneficial to remove the counter-force exerted by the distributional similarity of nouns in the input. But it is important to remember that the SRN is not explicitly trained to acquire

semantic categories; instead, it is tasked to predict sequences, which would not be very effective without knowledge of syntactic categories. This is a good way to think about how the goal of sequence prediction is not ideally suited for semantic category acquisition. Nonetheless, there is a way to reduce the impact of the counter-force that maintains superordinate (e.g syntactic) category structure: An incremental reduction in the distributional similarity of nouns across training. Hopefully, the point of figure 6.4 is now more clear. Only when training on AO-CHILDES in age-order does the distributional similarity of nouns decrease. That is, the counter-force that pushes nouns closer together, in the SRN's representational space, is gradually reduced. Training in reverse-age order has the opposite effect: Even though semantic differentiation is pushing probe words farther apart, the increasing distributional similarity of nouns should make it increasingly harder to do so. But why might it be better (for semantic category acquisition) to train on input with *decreasing*, rather than *increasing* distributional similarity of nouns, given that what matters is the *magnitude*, rather than the *direction of change* (across the input). I think this has to do with matching the state of the model to the state of the input across training time. During early stages of training the SRN is acquiring distinctions between syntactic categories, so it is useful to have a high distributional similarity of nouns. Nouns are more likely to cluster together in representational space, and this should speed acquisition of the category NOUN. When it is time to differentiate semantic categories within the category NOUN, the counter-force that maintains proximity between members of the category NOUN has decreased. The movement of words in representational space, due to semantic differentiation is now less constrained. Moreover, as more semantic distinctions are made, more free movement is allowed to the gradual reduction in the force that maintains proximity amongst members of category NOUN. On the contrary, during reverse age-ordered the training, the ideal

time for semantic differentiation, according to the pattern of distributional similarity shown in the left panel of figure 6.4, is at the very beginning of training. This is clearly a misalignment between the state of the model and the state of the input. Moreover, when semantic differentiation becomes possible during later stages of training, the distributional similarity of nouns is increasing, making it more difficult for movement of words in the category NOUN.

Interestingly, this decreasing trend was not observed for any other syntactic category investigated (verbs, adjectives, interjections, prepositions, conjunctions). This observation is consistent with the special role that the noun category plays in AO-CHILDES in facilitating semantic differentiation.

Note also the right panel of figure 6.4, which shows a performance trajectory related to how good the SRN is at predicting that a noun should occur next. The measure is based on the average precision, and is used here to quantify how much probability the SRN assigns to all nouns in the vocabulary when predicting the next word in a sequence that is actually followed by a noun. The actual measure is a composite measure as it is an average of all the average precision values obtained for all sequences ending with a noun. The results are consistent with my theory: The SRNs trained in age-order (in blue) assign more probability to nouns during the first half of training, and assign less probability to nouns in the second half of training compared to SRNs trained in reverse age-order. It can be said, that the SRNs trained in age-order initially treat nouns more as a unit than the SRNs trained in reverse age-order.

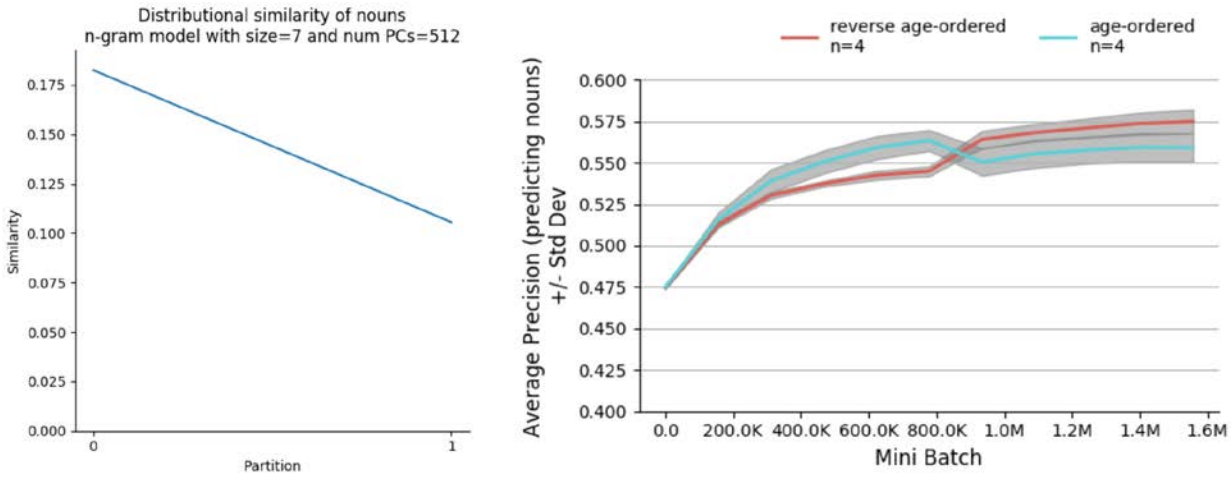

Figure 6.4
Left panel: Distributional similarity of nouns computed on partition 1 and partition 2. Noun representations were obtained by retrieving the rows corresponding to all nouns in the term-by-window co-occurrence matrix, and reducing their dimensionality to 512 via SVD. The average of pairwise cosine similarities between resulting representations is shown.
Right Panel: Mean-average-precision, computed on the SRN's output predictions given windows where the correct next word is a noun. A score of 1.0 would indicate the model is predicting only, and all nouns; a score of 0.0 would indicate the model is assigning zero probability to all nouns at the output layer.

Now, let's turn to a measure of differentiation in the model. As before, I have plotted cosine similarity of nouns, but this time, in the distributional space learned by the SRN. Figure 6.5 shows the trajectory of average pairwise cosine-similarities between probe words (top panels) and nouns (lower panels) as a function of training time (measured as the number of mini batches or update steps) averaged across 4 SRNs in each condition. Instead of using hidden state spaces (with sequences as input) to compute similarities, I used the input (left panels) and output (right panels) weights of the SRN to simplify interpretation. We can interpret distributional similarity here as providing an indirect measure of the counter-force that maintains the proximity of nouns in representational space. The similarity between noun representations at the end of training can be used as an indicator of how much semantic differentiation has occurred (higher

similarity means less differentiation). In all four panels, the end-of-training average similarity is larger for SRNs trained in reverse age-order (red) compared to SRNs trained in age-order (blue). This is in agreement with my theory, because it predicts greater semantic differentiation between probe words in the age-ordered training condition. The slightly greater similarity of the models trained in age-order during the first half of training shown in the top right and left panel, is also consistent with the theory because it predicts that nouns are initially represented closer in the SRN's representational space.

One thing to note is that similarity increases overall. Does this not violate the idea that semantic differentiation occurs, and that representations should become *less* similar over time? No, it only means that the representations of nouns are moving closer in representational space, as they should in a model trained to predict word sequences. Similarity should decrease only between probe words from different categories. In the analysis above, however, similarity is computed for all noun pairs.

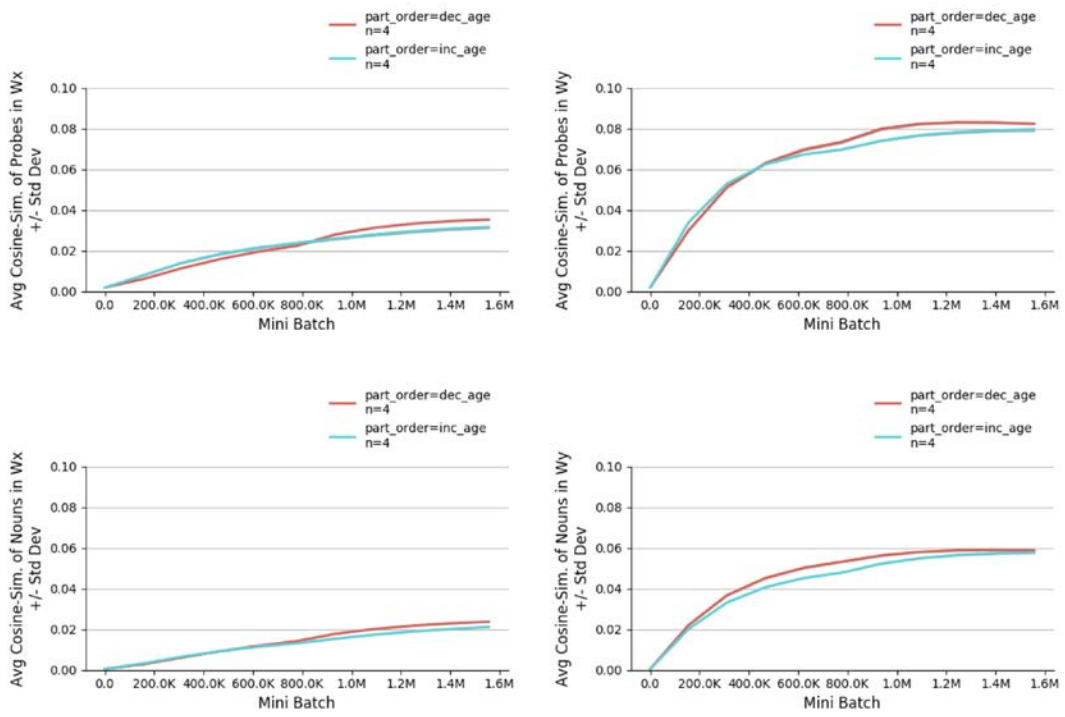

Figure 6.5
Average cosine similarity of input (left panels) and output (right panels) weights corresponding to probe words (top panels) and nouns (lower panels) obtained at consecutive intervals during training. Red, SRN trained in reverse age-order; blue, SRN trained in age-order. Each point on the line represents the average, across 4 simulations, of average cosine similarities.

Before concluding this section, I provide an illustration of what is occuring at the word-level, rather than at the category-level. This demonstration is not directly testing a component of my theory, but provides a very straightforward example of what semantic differentiation looks like under ideal and not-so ideal conditions. Specifically, of interest is the case in which the movement of a word's representation through the SRN's representational space is influenced by both its semantic category and its superordinate syntactic category. In figure 6.6, I plotted the similarity (pearson-correlation) between hidden state representations for *dog, cup, mom, three, and june*. Hidden state representations for each word were obtained by averaging a trained SRN's hidden states for windows in which the given word occupied the last position. This is the same pipeline that was used to obtain word representations in the experiments described in

chapters 2 and 3, and was used here for convenience; there is no reason why input or output based representations should provide qualitatively different results. Each panel tracks the similarity of the representation of the word indicated by the panel title with representations of other words, indicated by the figure legend. In the first four panels semantic differentiation clearly pushes representations of related words closer together (*cat* and *dog* in the top left panel, *cup* and *plate* in top right panel), and clearly pushes representations of unrelated words farther apart (*dog* and *february* in top left panel, *cup* and *february* in top right panel). This is business as usual; however, the value of this demonstration comes from inspecting what happens to the similarity of *june* to all other words. No clear pattern of semantic differentiation is apparent, as its representation remains approximately equally similar to all other six words. Ideally, *june* should become more similar to *february* as both refer to months in the year. The reason this does not happen is because *june* is not exclusively used to refer to the month in AO-CHILDES; instead it is frequently used to refer to a person of the same name. The lack of semantic differentiation is a great example of what happens when there are two counteracting forces pushing a word's representation in two different directions. One the one hand, june is pushed towards the space occupied by members of the category MONTHS, but on the other hand, its movement is constrained by being a member of the category PROPER NOUN. In the end, *june* remains undifferentiated: Its position in representational space represents some average between members in the category MONTHS and members in the category PROPER NOUN. This example provides a clear demonstration of what movement through representational space looks like when semantic and syntactic membership are in conflict.

This example is an extreme case in which syntactic and semantic category membership are actually mutually exclusive, and therefore should not be confused with the more frequent

scenario in which a words syntactic membership does not conflict with its semantic category membership. For example, *cat* is both a noun, and a member of the category MAMMALS. When cat is used, both categories are applicable; however, when *june* is used in AO-CHILDES it cannot simultaneously refer to both a person and a month. As noted previously, this demonstration was provided to help the reader gain a better understanding of semantic differentiation in the SRN, rather than to directly test a component of my theory.

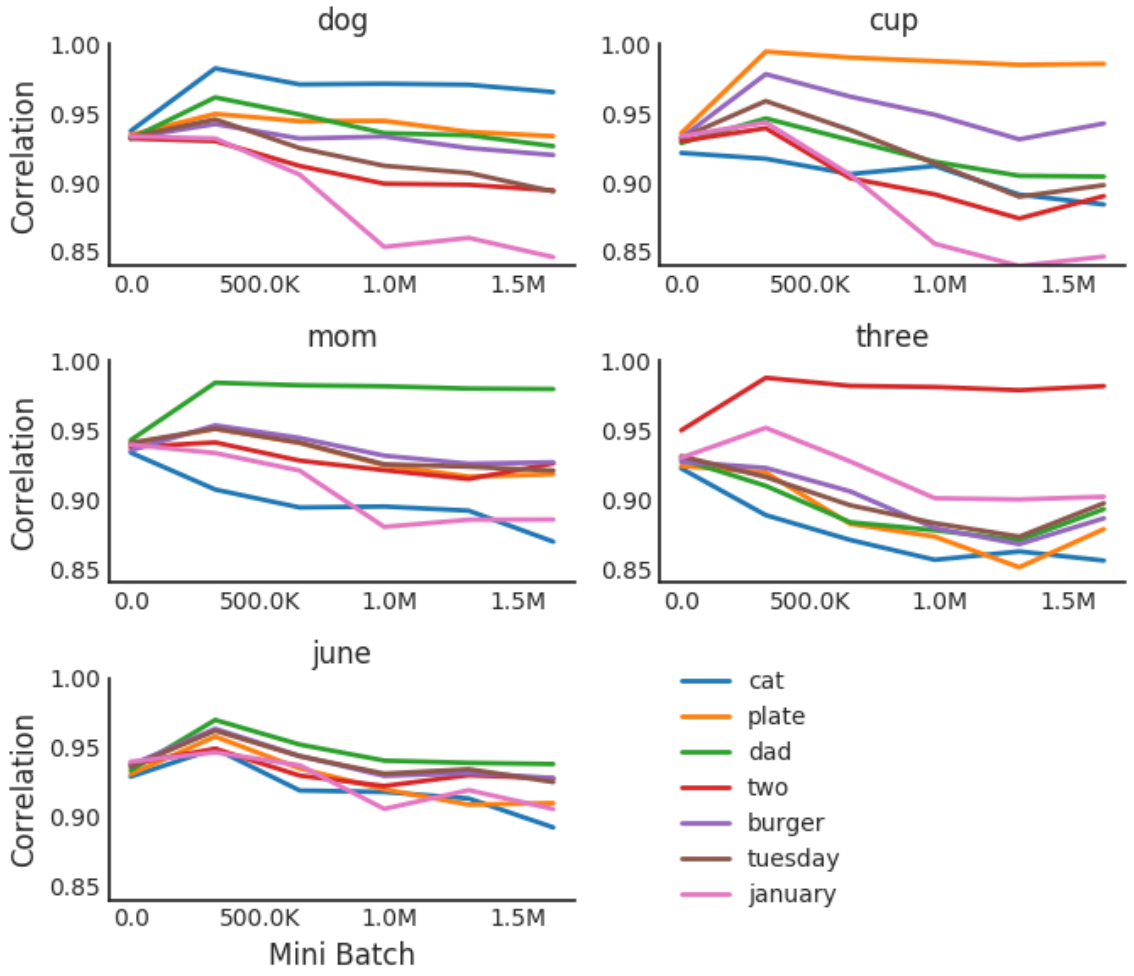

Figure 6.6
Pearson-correlation between average hidden layer representations of 7-word sequences in which either *cat, plate, dad, two, burger, tuesday, or january* was in the last position in the window, across training time (in number of weight updates). Correlation between unrelated words drops across training time. For example, the correlation of *dog* with all other seven words is shown in the top left panel. Note that unrelated words continue to move away from *dog* (in hidden state space), while the related *cat* has moved closer to dog. This pattern of differentiation is not detected for *june*, which is used both to refer to the month and a person in the AO-CHILDES corpus.

The special role of nouns

The theory developed here ascribes a special role to the incrementally changing distributional properties of nouns in giving rise to the age-order effect. Specifically, it is argued that the age-order effect is a consequence of the relative ease of acquiring the category during

early training. What evidence exists to support this claim? I conducted a comparison of the hidden-state spaces at the end of training of models in various training conditions. In all simulations the models were trained on 256 partitions of AO-CHILDES. I trained 5 SRNs in age-order, and 5 SRNs in each additional training condition, in which the order of presentation of training examples was varied systematically. I ordered the partitions by increasing entropy of words that are left-adjacent to any member of a specific category. For example, words that are left-adjacent to nouns in the sentence *the man gave his uncle a gift* are *the* and *his*. For each partition, the left-adjacent context works are collected, and the entropy of their distribution is computed. The partition associated with the lowest entropy is presented first, and subsequent partitions are presented in the order of increasing entropy. The categories used in this analysis are mostly syntactic (e.g. conjunctions, nouns, prepositions, verbs), but I included punctuation, which includes various utterance boundary markers such as periods) and a category consisting of all 532 probe words. The entropy was chosen as a way to quantity the ease with which a category can be acquired. If the entropy of left-adjacent words is low, then the category should be easier to learn compared to if the entropy is high. In the former case, any left-adjacent word contains less information about the identity of the word that comes after it. If it is true that the ease with which nouns can be predicted during early training plays a special role during age-ordered training, then an ordering of partitions based on noun contexts which are most predictive of nouns during early training, should produce a hidden-state space (computed over all vocabulary words) that is most similar to models trained in age-order compared to models trained in alternative training conditions. The results of this analysis are shown in figure 6.7. Each row in the heatmap represents a model on one of 7 training conditions indicated by the y-axis label. For example, the first five rows indicate models which were trained on 256 partitions ordered by the

age of the target child. The next five rows represent models trained on 256 partitions ordered by the entropy of words which are left-adjacent to conjunctions. The order in the rows is preserved across the columns. Each cell is the average of the pairwise cosine similarities between hidden states for the same vocabulary word retrieved from a model indicated by the row and another model indicated by the column. As such, each cell represents a rough indicator the similarity of the semantic spaces obtained by two different SRNs. The important comparisons are in the first five columns, which indicate how similar a model's semantic space is to the semantic space of models trained in age-order. The semantic spaces of models trained on partitions ordered by increasing entropy of noun contexts are most similar to models trained in age-order, as indicated by the darker red color in the first 5 entries of rows 10-15. A dark red represents higher similarity, whereas yellow and green indicate lower similarity. The similarity between models trained in age-order and models trained in any other order are less similar. This means that ordering partitions by a measure related to the difficulty of predicting nouns results in a pattern of learning that is most similar to that of models trained in age-order.

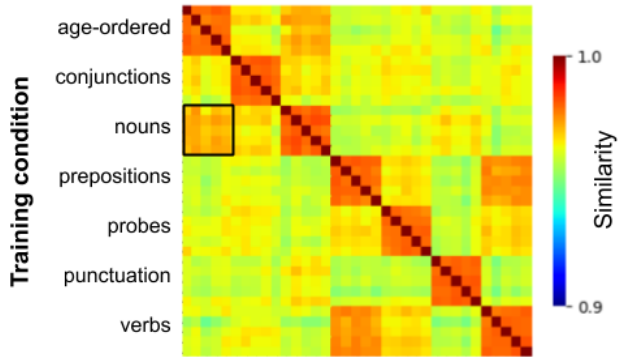

Figure 6.7
Correlation matrix comparing the hidden state spaces of models trained on 256 partitions ordered in age-order (first 5 rows and columns) and models trained on 256 partitions ordered by increasing entropy of words left-adjacent to members of the category indicated by the row label.

Examples

What exactly do I mean when I speak of the distributional similarity of nouns? And how exactly does it influence semantic category learning? To clarify, I provided concrete examples below. I did not pick them from AO-CHILDES, so they should not be taken at face value. There are numerous was in which noun contexts can be made more or less similar, and the following two examples represent only a small fraction. The first three utterances (a), (b), and (c), below, each share the construction *Where did that _*, and in each case, the slot is occupied by a noun. Thus, the noun context is the construction *Where did that*. Because all three nouns occur in the same context, the distributional similarity of nouns can be said to be extremely high. This state of affairs would enable to the SRN to quickly acquire a category for nouns. In this example, *bus*, *car*, and *dog*, would all be represented by the same vector, because each was seen under identical circumstances. This is the 'good start' that I have been talking about. Semantic differentiation is only informative if the starting relationships between word representations are not influenced by irrelevant information. In this case, there is no irrelevant information. However, in a syntactically more complex language environment, such as partition 2, the utterance (d) is more likely to occur. It makes use of the familiar construction *Where did that _* in a way that violates our expectations about how it should be used. Concretely, the variable slot was previously always filled by a noun, but in utterance (c), the slot is occupied by a verb. While grammatically correct, utterance (c) makes it more difficult to learn a category that consists entirely of nouns. After all, a distributional analysis would suggest that come should be treated like *bus*, *car*, and *dog*, which are all nouns. In other words, the distributional similarity of nouns has been lessened. The context which was previously perfectly diagnostic of a noun is no longer a perfect indicator. If the SRN were to begin training on input in which utterances (a) - (d) were

to occur, it would have a harder time forming category that consists entirely out of nouns. This can have negative consequences on subsequent semantic differentiation, because the movement of noun words in representational space would be influenced by the movement of words that the SRN thinks are noun-like, but are in fact not nouns (e.g. *come*).

*(a) Where did that **bus** go?*
*(b) Where did that **car** go?*
*(c) Where did that **dog** go?*
*(d) Where did that **come** from?*

Relationship to the good-start hypothesis

How does the theory developed in this chapter relate back to the good-start hypothesis that was developed in chapter 5? The discussion about progressive differentiation, and singular values has largely clouded the straightforward notion that speech to younger children can provide a better starting point for learning semantic categories compared to speech to older children. Let's return to the question asked in the beginning of the chapter: what is a 'good start'? The answer has to do with the syntactic complexity of the input. Specifically, it appears that the prominence of the noun category relative to other syntactic categories in the input plays a special role in facilitating subsequent semantic differentiation of probe words (which are all nouns). A clear separation of nouns from other words in the SRN's representational space during early training seems to be advantageous for subsequent semantic differentiation. It is easiest to understand why this is the case by imagining that members of different syntactic categories initially occupy a distinct area in the SRN's representational space. Over the course of training, the finer-grained differences that exist between members of each category in the input will be reflected in the movement of their representations within the representational space initially defined by the category at the start of training. The dimensions that are important for

distinguishing between members of one syntactic category should ideally not be the same dimensions along which members of a different syntactic category are differentiating, because syntactic categories represent non-overlapping sets of words with very distinct linguistic functions. Therefore, the dimensions in the representational space that are most important for semantic categorization in this work are distinctions that are unique to nouns. Achieving maximum performance requires that probe word representations do not differ along any dimensions which are not unique to the noun category. If so, this would mean that the dimensions that are important for distinguishing between members of non-noun syntactic categories have influenced the differentiation of probe words. I understand 'good start' as minimizing the chances that this blurring or overlapping of dimensions occurs. It does this by establishing - early during training - a space within which semantic differentiation can occur that is as far away as possible from the spaces occupied by other syntactic categories. Thus, over the course of training, the likelihood that a meeting, in representational space, between members of *different* syntactic categories occurs is minimized. By so doing, the dimensions important for distinguishing non-probe words do not interfere with those that are important for distinguishing between probe words. For example, the distributional distinction between ANIMALS and BIRD should not be blurred by distributional differences between, say, *running* and *climbing.*

Limitations

While the mathematical connection between an increased number of unique constructions and smaller singular values is strong (argument 2 and 3), the consequences of smaller singular values on semantic category learning in the SRN is more speculative (arguments 4 and 5). Saxe, McClelland & Ganguli showed that a singular dimension associated with a smaller singular

value is acquired more slowly by deep linear networks. While I could have stopped here, and conclude that semantic differentiation that takes place during reverse age-ordered training simply takes more time, I added an additional, more speculative component: Because acquisition periods of distinct dimensions of the input are slower during reverse age-ordered training, they are more likely to overlap. Thus, movement in the SRN's representational space is at greater risk of being influenced by several (possibly competing) adjustments to the representational space. Why did I not stop at the simpler explanation? While I cannot rule out this simpler alternative, I think it is not sufficient to explain the age-order effect. In fact, semantic differentiation should be slower, but I don't think that that is the full story. After all, I have shown in the previous chapter that increasing training time on partition 2 and reducing training time on partition 1 does not eliminate the age-order effect. If it was simply the matter of time, then additional training iterations should have equalized the gap in semantic categorization performance between the two training conditions. Given this result I proposed early on that a good theory must consider explicitly the effect of partition *order*, rather than exclusively on their *complexity*. I think that training in age-order results in a qualitatively different organization of the SRN's representational space, and that this organization results in less rearrangement when crossing from partition 1 to 2, than the SRN trained in reverse age-order. If the age-order effect was only due to a difference in complexity of the acquired representational space, rather than some (additional) qualitative difference, then it would be possible to eliminate any performance differences by balancing the amount of training time allotted to each partition. Having failed to find such a balance, I am forced to consider an alternative theory to explain this discrepancy.

Another important limitation concerns the sequential nature of the regularities that the SRN used as its basis for learning semantic category structure. My theory does not take into

consideration the role that distance plays in influencing the semantic dependencies that are acquired by the SRN. Having computed balanced accuracy on bag-of-word representations for partition 1 and 2 windows of various sizes (1 to 7), I found that information about semantic category membership (the categories used in this work) drops off more quickly with distance in partition 1 compared to partition 2. This means that there are more semantic dependencies in partition 2 that span longer distances compared to partition 1. This is most likely related to the fact that utterances are, on average, longer in partition 2. Semantic dependencies within utterances are more likely to be separated by a larger number of words in partition 2 simply because utterance length is larger. Therefore it is possible that the same semantic dependency in partition 2 may span a shorter distance in partition 2, where utterances are shorter. Because longer distance dependencies are more difficult for the SRN to learn (Hochreiter & Schmidhuber, 1997), this could influence semantic category acquisition. For example, the SRN trained on partition 1 first, where semantic category information is primarily found in dependencies spanning shorter distances, may be more prepared to recognize similar semantic dependencies in partition 1 where semantic dependencies span greater distances.

A third limitation is that the theory does not take into consideration the role that weight entrenchment might play in explaining the age-order effect. Weight entrenchment is the gradual reduction in plasticity over the course of learning in nonlinear neural networks (of which the SRN is a member of), and has been attributed to a gradual reduction in the effectiveness of backpropagation when using a nonlinear activation function (see Munro, 1986, for an early discussion of this phenomenon). Because weights are typically set to random values close to zero at the start of training, weights are most responsive to learning at the initial stages of learning. Consequently, earlier-learned patterns may become entrenched in the weights. A sure way to

enforce weight entrenchment is to use and adaptive optimization procedure, such as AdaGrad to replace the vanilla SGD [14]algorithm at the heart of backpropagation-based learning. Briefly, AdaGrad reduces the magnitude of updates to parameters in proportion to the sum of the magnitudes of previous updates to the same parameter. Overall, this has the effect of gradually reducing the impact of newly learned patterns on existing patterns acquired by the model. If the patterns obtained via training on partition 1 are more generalizable to partition 2 than vice versa, weight entrenchment would elegantly explain why training in age-order would result in less interference compared to training in reverse age-order. When trained in reverse age-order, the SRN has encoded sequential regularities during training on partition 2 which do not generalize well. Conversely knowledge acquired during training on partition 1 would generalize to partition 2, reducing the potential impact of interference on previously acquired knowledge. Because the SRN trained in age-order is in less danger of losing previously acquired knowledge, end-of-training performance is greater than that of the SRN trained in reverse age-order. While this explanation is appealing, I have shown that the age-order effect is at least resistant to the choice of optimization procedure (SGD vs. AdaGrad).

Lastly, I have assumed throughout this chapter that the learning dynamics of the deep nonlinear SRN used in this work resembles the learning dynamics of deep linear networks studied by Saxe, McClelland & Ganguli (2019). To my knowledge, their findings have not been replicated in nonlinear networks, and nor shall I attempt to do so. However, in the next chapter, I will explicitly test whether superordinate category distinctions are learned before subordinate category distinctions. If so, this would be strong evidence that progressive differentiation occurs in the SRN, and that the theory of Saxe, McClelland & Ganguli (2019) also applies to the SRN.

[14] Stochastic Gradient Descent

CHAPTER 7: SIMULATIONS WITH ARTIFICIAL INPUT

In the previous chapter, I laid out a theory to explain the facilitatory effect of age-ordered training on the ability of the SRN to acquire knowledge of the semantic category of words. In this chapter I describe two experiments designed to test this theory. To provide a brief overview of what is to follow: First, I developed a simple artificial input, consisting of 2-element sequences with hierarchical semantic category structure. In the first part of the chapter I describe how I used the artificial input to probe the SRN's learning dynamics. Specifically, I asked whether the SRN undergoes progressive differentiation, as formalized by Saxe, McClelland & Ganguli in their theory about deep linear networks. In the second half of the chapter, I add simple syntactic distinctions to the structure of the artificial input. I asked whether an incremental change in the complexity of the surface structure of syntactic, but *not* semantic categories, can affect acquisition of the semantic category structure. Interaction between syntactic and semantic category learning is a key component of the theory developed in chapter 6. Similar to the procedure used in chapter 3 to show the age-order effect, I trained a group of SRNs on artificial input with input ordered either by increasing or decreasing complexity, and compare their semantic categorization performance trajectory.

Generating Toy Input

The first set of simulations uses the following artificial input: 5 million sequences of 2 items each. I will refer to the first item as the probe word, and the second as the context word. The tasks of the SRN is to predict the context word from the probe word. There are 1024 probe words and 1024 context words. Each probe word belongs to one of 32 categories. It might be

helpful to think of these categories as semantic categories, to connect the ideas develop here back to experiments conducted with AO-CHILDES in chapters 2 and 3. Both the number of sequences and the number of categories were chosen to match as closely as possible the conditions under which the age-order effect was observed.

The reason that a context word follows a probe-word is simple: The SRN's internal representation for a probe word is updated by backpropagating the error between the SRN's prediction about the next word and the correct next word. Simply put, if the context word was positioned first in the sequence, then the internal representations for probe words would never change. We are interested, of course, in the SRN's internal representations of probe words, because they are the words that are given category structure in the input, and the category structure of probe words is defined in terms of context words. Specifically, the category membership of a probe word is defined in terms of its co-occurrence distribution over context words. Probe words in the same category have co-occurrence distributions over context words which are more similar to each other than to any other probe word that is not in the same category. While predicting context words, the SRN is implicitly learning the distribution of co-occurrences associated with each probe word, and representing it in the input weights (also known as the embedding layer). At the end of training, the input weights come to represent the knowledge which context words are likely to follow each probe word, and therefore encode category membership - possibly, at multiple levels in some hierarchical category structure.

How are probes assigned to their categories? Because I needed not only the ability to programmatically assign single categories to probes, but also to superordinate categories in a hierarchically organized tree, I turned to probability theory. A routine, known as the 'branching diffusion process' generates binary vectors (+1 and -1) with probabilistic hierarchical structure. I

used a variant of the branching diffusion process which derives hierarchical structure from a binary tree (a hierarchical tree where each node has exactly two branches connecting it to its subordinate nodes). Figure 7. 1 provides an illustration of how it works: The process starts with a vector of size 1, containing only the number +1. Next, this vector undergoes a number of expansions, in which each element is copied twice. Thus, if there are 10 expansions, the resulting vector is of size 1024, containing only +1s. The key insight to constructing a vector with probabilistic hierarchical structure is to allow copy errors, or mutations. For all experiments, I used a mutation probability of 0.01, meaning that at each 'copy step', the probability that a sign change (from +1 to -1, and vice versa) occurs is 0.01. Because this probability is relatively low, any sign change at any level in the conceptual binary tree (of copy steps) will be apparent in the resulting vector. Most importantly, a sign change, for example, somewhere at mid-level in the binary tree, only affects nodes below it. Due to this hierarchical relationship between nodes in the branching process, the output vector approximates the hierarchical structure of the binary tree (of copy steps). Put differently, the probabilistic relationship between two elements in the vector is determined by the distance between nodes in the conceptual binary tree. Using this routine, I generated a vector for each context word. The resulting 1024 vectors, each containing 1024 values, represent the complete sequential structure of the input. Specifically, each context word's vector represents which probe word is allowed to precede it. Each index in the vector is associated with a unique probe word, and a +1 means that the probe word associated with the index is allowed to precede the context word, and a -1 disallows this.

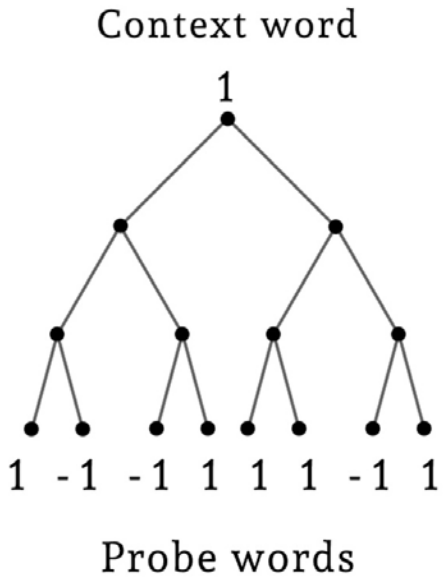

Figure 7.1
A schematic of the branching diffusion process used to generate sequential regularities in the toy data. Each branch in the binary tree represents an opportunity where a sign change can occur (from -1 to +1 or +1 to -). The probability of this change was set to 0.01 for all experiments. The resulting vector of +1s and -1s was used to define sequential regularities in the toy input. Because underlying each vector is the same hierarchical structure, the same structure is reflected in the sequential regularities, and therefore the distributional cues in the toy input.

I generated a vector of +1s and -1s for each context word, but the question of how probes are assigned to categories remains. Most of the work has already been done. But before moving on, let's define a new concept. I will refer to the matrix, that results when horizontally concatenating all the row vectors as **L** (short for 'legal transitions'). **L** is of shape [1024, 1024], where rows represent context words, and columns represent probe words. It is now relatively straightforward to understand where the sequential regularities underlying the toy input originate. Essentially, anywhere there is a +1 in **L**, the context word associated with the row index is allowed (legal) to follow the probe word associated with the column index. On the flipside, anywhere there is a -1 in **L**, co-occurrence between the context word and probe word is

disallowed (illegal). In fact, this matrix a term-by-window co-occurrence matrix which we have already encountered in the previous chapter, except that 1-word windows are used here because each sequence consists of two words. With the sequential structure of the input defined, we can compute category membership. Notice, that category membership is computed after **L** has been generated. This ensures that the hierarchical structure in **L** is preserved in the category assignments. The first step is to use a simple hierarchical clustering algorithm (using the Python routine *scipy.cluster.hierarchy*) with the correlation matrix of **L** as input, to retrieve the (approximate) hierarchical structure implicit in **L**. To illustrate the resultant structure, I used the structure retrieved by the clustering algorithm to cluster the rows and words of the correlation matrix, and plotted it, shown in figure 7.2. Each row represents the distributional similarity of a context word (represented by the row) with a different context word (represented by the column). Values farther away from dark blue indicates greater distributional similarity. Each light square in the figure represents a category at some level in the hierarchy. These squares become increasingly brighter the smaller their size, until, eventually, squares are the size of a single row and column. The red diagonal (similarity = 1.0) indicates that each word is perfectly correlated with itself.

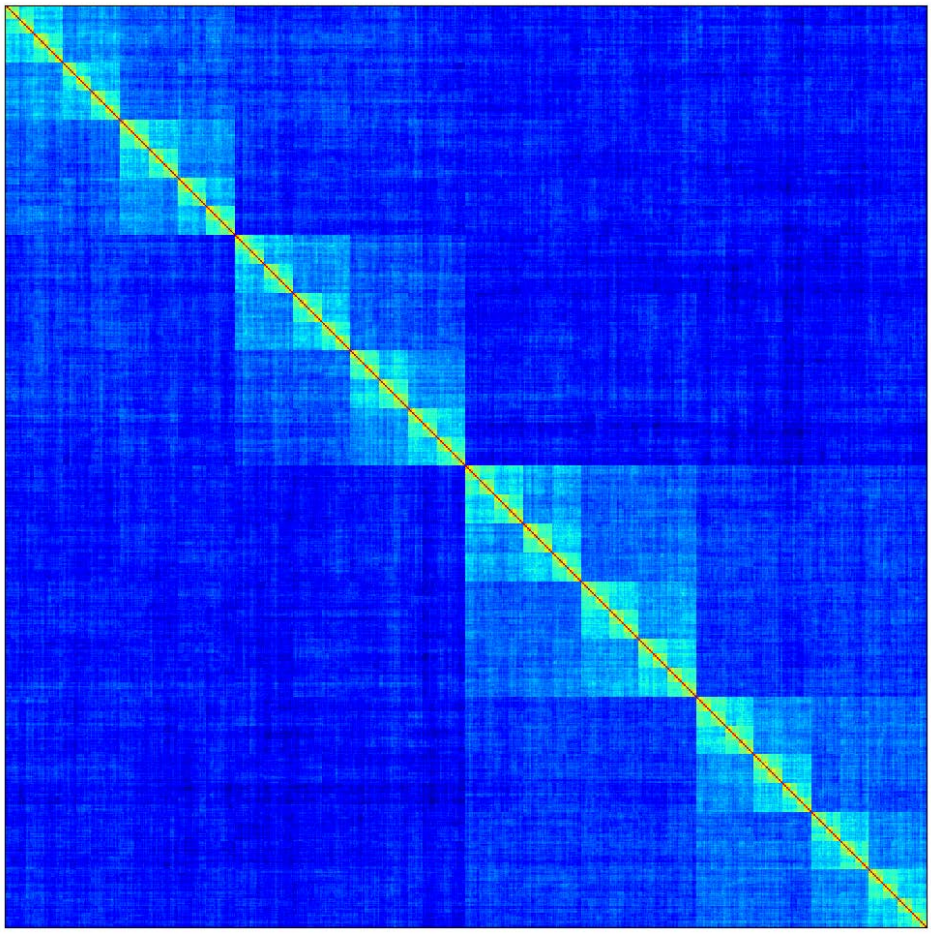

Figure 7.2
The correlation matrix of **L**, clustered along rows and columns to illustrate the hierarchically organized subcategory structure. The number of rows is identical to the number of columns and reflects the number rows in **L**. Values indicate the cosine similarity between the co-occurrence pattern of the context word represented by the row with a different context word represented by the column. Dark blue = 0.0, dark red = 1.0.

I used the results of the hierarchical clustering to assign each probe to a category at each level in the hierarchy. The result is a subcategory structure that is hierarchically organized, where each probe is a member of one category at each level in the hierarchy. Whatever hierarchical structure was implicit in **L**, it is now explicit in the category assignments, and these will be used to evaluate categorization performance.

At this point it is important to note that the artificial input does not have any syntactic categories. Because I conceive of the probe words as members of semantic categories, all of the sequential regularities specified by **L** may be best thought of as signaling semantic, rather than syntactic category membership. It might therefore be useful to think of probe words as nouns, and as the artificial input as consisting of only a single syntactic category, rendering it syntactically 'empty'. If I wanted to add a new syntactic category, representing, say verbs, I would also have to add an additional set of context words which are used exclusively in combination with the verbs. If I continued to use the same context words assigned to the nouns, the 'verbs' would be indistinguishable from the nouns, and this would violate the definition of a syntactic category. In the second set of simulations, described in the second half of this chapter, syntactic categories were added in this fashion. One reason for keeping the toy input as simple as possible in this first set of simulations, is to be able to learn more about the 'default' learning dynamics of the SRN. This can be achieved when all the idiosyncrasies of the input have been stripped away. The simple assignment of probe words to their semantic categories allowed me to neatly and evenly divide the items in the artificial input into categories at multiple levels in a hierarchy. A more complex input with additional syntactic category structure would have made this needlessly difficult. Being able to assign the probe words into categories at multiple levels was crucial for testing my assumption that 'progressive differentiation' actually occurs in the SRN. In fact, I used a branching diffusion routine similar to the one used by Saxe, McClelland & Ganguli to develop their mathematical description of the learning dynamics of deep linear networks.

Simulations 1: Progressive Differentiation

To test whether the SRN undergoes progressive differentiation, I first generated a single corpus by sampling randomly from the set of sequences that are legal in **L** (entry associated with probe word and context word is +1, not -1). As stated before, I sampled 5 million sequences to keep the size of the corpus consistent with AO-CHILDES. While consistency with AO-CHILDES is strictly not necessary in this first set of simulations, it is an important factor in the second set of simulations. I trained at least two SRNs with a hidden layer size of 128 in four conditions, varying both the learning rate and the optimizer. During training, I tracked the balanced accuracy associated with the different levels in the hierarchically organized subcategory structure. There are 5 balanced accuracy trajectories in total, each associated with a level in the subcategory structure: the first level (with 2 distinct categories), the second (with 4 distinct categories), the third (with 8 distinct categories), the fourth (with 16 distinct categories), and the fifth level (with 32 distinct categories). In total each SRN iterated 40 times, over 5 million sequences (seen in batches of 64). No incremental training regime was used, nor was there any incremental structure in the input. If the SRN undergoes progressive differentiation of the semantic category structure in the artificial input, then balanced accuracy associated with the distinction between the 2 top-most categories should peak first, and balanced accuracies for subsequent lower-level distinctions should peak in order of their level in the category hierarchy. The results are shown in Figure 7.3. Each panel shows semantic categorization performance (as measured by the balanced accuracy) computed for the 5 different levels at equally spaced intervals during training. The top panels show balanced accuracy trajectories averaged over 12 and 8 SRNs, trained with SGD and learning rates 0.06 (left panel) and 0.3 (right panel). In both cases, the hallmark of progressive differentiation is clearly visible. Balanced accuracy associated

with the first level in the semantic category hierarchy peaks first, and consecutive peaks are associated with semantic category distinctions at progressively lower levels in the semantic category hierarchy (red=1st level, yellow=2nd level, green=3rd level, blue=4th level, purple = 5th level). Interestingly, the first peak in balanced accuracy (1st level) is followed by a dramatic decline, and a gradual leveling out. This pattern is increasingly less distinct for balanced accuracy trajectories peaking later. In fact, the balanced accuracy associated with category distinctions at the 5th level does not decline when the learning rate is relatively low (0.06, left panel) and declines only modestly when the learning rate is larger (0.03, right panel). One interpretation is that as the SRN is learning lower-level descriptions of the input (associated with distinctions at progressively lower levels in the category hierarchy), it is no longer useful to hold on to a higher-level description. Acquisition of knowledge at a lower level enables greater sequence prediction performance, because the SRN is able to make finer-grained distinctions.

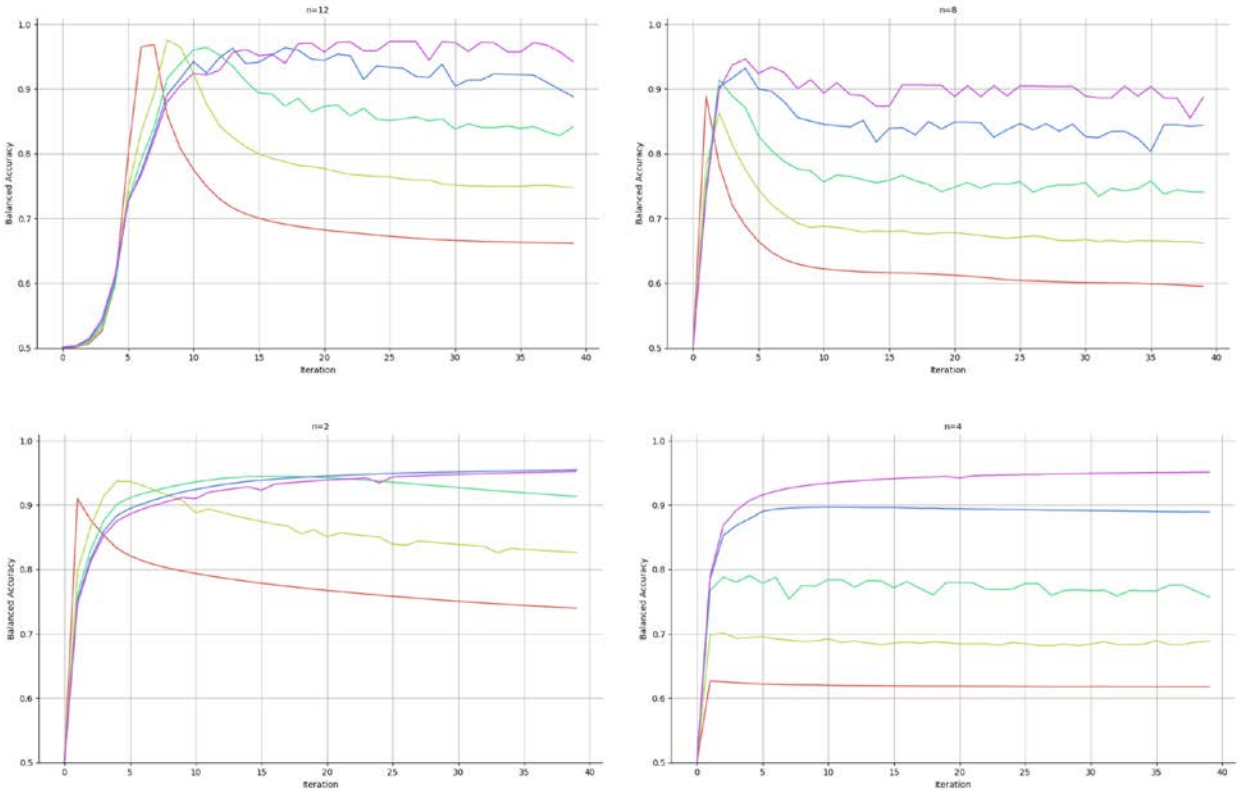

Figure 7.3
Balanced accuracy at consecutive intervals during SRN training with SGD (top panels) and AdaGrad (lower panels) on artificial input. Learning rates in clockwise order starting at the top left: 0.06, 0.3, 0.006, 0.003. Color indicates which level in the category hierarchy is evaluated. Red=1st level, yellow=2nd level, green=3rd level, blue=4th level, purple = 5th level

What about the SRNs trained with AdaGrad (performance shown in lower panels)? The mathematical description provided by Saxe, McClelland & Ganguli were obtained by training and simulating networks using SGD, the simplest implementation of backpropagation-based optimization. An important open question is whether similar learning dynamics are observed when the optimizer is not SGD. While this is an important question in its own right, is important to test it here because the SRNs for which the age-order effect was observed were trained with AdaGrad, and not SGD. I have said previously that the age-order effect does not strictly rely on AdaGrad, but replacement of AdaGrad with SGD reduces overall semantic categorization

performance, and does reduce the age-order effect. The lower panels of figure 7.1 tell a complex story: While the pattern of balanced accuracy trajectories shown in the left panel is in agreement with progressive differentiation, no distinct peaks in performance are detectable in the right panel, which shows results of SRNs trained with a larger learning rate (0.006 vs. 0.03). Does this mean that a larger learning rate or optimization other than SGD can result in learning dynamics other than progressive differentiation? (If so, what else are these networks doing?) While this is difficult to imagine, it is a possibility. A more plausible alternative is that progressive differentiation does occur, but is obscured by the fast learning rate. The model trained with a learning rate of 0.03 converges very quickly, and it is possible that progressive differentiation of the first 5 levels in the category hierarchy occurred so early during training that the evaluation intervals were too large to detect it. Another possibility is that method for detecting progressive differentiation is not sensitive enough. Note, that the balanced accuracy is only one of many ways to characterize category learning in the SRN, and it is at best an indirect measure of what is actually happening to the SRN's representational space over time. Saxe, McClelland & Ganguli did not use the balanced accuracy as I did; instead, they derived the learning dynamics from first principles, and verified their results by comparing predicted trajectories of 'effective singular values' (singular values derived from the network) to trajectories actually observed during training. Nonetheless, the results are convincing that progressive differentiation does occur in the SRN, even if it is not clearly detectable in all circumstances.

Adding Syntactic Categories

In this section, I test another assumption underlying my theory: Does a change in the distributional properties of non-probe words influence acquisition of the semantic categories of

probes words? The kind of change I have in mind is a change in the complexity of the surface structure, which I have used synonymously with the number of unique constructions present in the input. Notice, that varying the number of unique constructions overall is not a direct modification of the semantic category structure of the input. Increasing or decreasing the number of unique constructions may affect sequences in which probe words occur, but does not change the underlying semantic category structure of the input. Instead, such a change would affect all words in the vocabulary equally. To be precise, this is true only if such a change was implemented artificially as is done here. As we have seen before in AO-CHILDES, the number of unique constructions is correlated with numerous other variables, such as density of various grammatical categories, MLU, and age of the target child. The benefit of using artificial input, is of course that I can vary complexity globally and be certain that probe and non-probe words were affected equally, and without possibly introducing additional confounding trends. In fact, I have chosen to vary the complexity of the artificial input independently of probes, meaning that sequences in which probe words occur are left entirely unchanged. Showing that such a change can still impact acquisition of semantic categories of probe words is a stronger test of the theory than if probe words were in any way affected by such a change.

To add syntactic category structure, I simply extended the existing matrix, **L**, which specifies the legal sequences. I added a different number of rows and words and partitioned, depending on the number of syntactic categories I wish to add. I tested 3 scenarios, in which either 1, 2, or 3 equally sized syntactic categories are added to **L**. Each syntactic category consists of words (represented by the rows, as before) which are grouped together by virtue of occurring with the same probe words (represented in the columns, as before). The context words that define one syntactic category are allowed to occur only after members of the syntactic

category; they are not shared between syntactic categories. This is an extreme view on the definition of a syntactic category, and was chosen only to keep evaluation as simple as possible. What do the co-occurrences within syntactic categories look like? To keep evaluation simple, and because my theory is agnostic about the category structure in general, I have opted to randomly assign each member to a context word with some fixed probability. Assignments occurred independently of each other. The probability that a member is allowed to occur with a context word is the same probability that a probe word is allowed to occur with a context word. Notice that I did not create hierarchical structure within syntactic categories. Hierarchical structure within probes allows evaluation of categorization performance at different levels, and because I am not interested in acquisition of non-probe category structure, I opted for random co-occurrence structure within syntactic categories Notice that this results in syntactic categories that are not qualitatively different; in natural language this may not be the case as, say, adjectives and nouns may have differing within-category structure. I do distinguish between nouns (probe words) and all other syntactic categories, because the within-category structure of nouns is the only one which is hierarchically organized. Even that distinction, however, is not necessary, to test my theory; the fact that the category structure of probe words is organized hierarchically is simply a holdover from the previous set of simulations in which I was interested in the different rate at which the SRN acquires category distinctions at each level.

The most important variable (in my theory) is the probability that a non-probe word is allowed to occur with a context word, henceforth referred to as P_1. P_1 is initially set to the probability that a probe word occurs with a context word, but importantly, can be varied independently. The probability that a probe word occurs with a context word, is a function of the mutation probability governing the branching diffusion process that generated the subcategory

structure underlying the hierarchical organization of probe word categories. Given a mutation probability of 0.01, this probability is approximately 0.92 (with slight variation if not reusing the same random seed). Because my theory predicts that varying the total number of legal sequences involving non-probe words can influence acquisition of probe-word categories, and because P_1 directly influences this number, P_1 can be viewed as the independent variable (and balanced accuracy is the dependent variable). But P_1 is not manipulated directly because this would cause trouble when trying to construct a corpus with incremental structure. To create a corpus, I must first create **L**, and then sample from **L** (5 million times to create a corpus of 5 million sequences. To vary the complexity of the corpus (the number of legal sequences, which is proportional to P_1) incrementally by varying P_1 directly, I would have to create another matrix **L**. But, because values in **L** are determined randomly, stitching together two corpora sampled from two different L matrices, would not result in incremental structure. Instead, the structural change would be total, not incremental. Put differently, structure that existed in the first corpus is not preserved in the second corpus, or vice versa. To remedy this issue, I fix P_1 at 0.94 in all simulations, and only vary the probability that a +1 in **L** is actually considered when sequences are sampled from **L**. I will refer to this probability as P_{legal}.

I generated three corpora, each with a different number of syntactic categories. So far, each row in **L** represents a probe word. Because my theory is agnostic to the precise number of syntactic categories (as long as there are some), I generated corpora with 1, 2, and 3 syntactic categories. Each additional category consists of 512 words, and 512 context words. The size of syntactic categories was made constant for simplicity, and such that the vocabulary size of the largest corpus is identical to the vocabulary size used to train on AO-CHILDES. I will refer to

the resulting corpora as C_x with subscript x indicating the number of added syntactic categories. The vocabulary size of each corpus is calculated below.

- vocabulary size of $C_1 = 512 + 512 + (512 + 512) \times 1 = 2{,}048$
- vocabulary size of $C_2 = 512 + 512 + (512 + 512) \times 2 = 3{,}072$
- vocabulary size of $C_3 = 512 + 512 + (512 + 512) \times 3 = 4096$

As an aside: While there are, strictly speaking, 10 million tokens in each corpus, the SRN is actually only fed 5 million tokens, because each second token is only used to compute the error at the output layer. In fact, the artificial input is best not understood as a single sequence of tokens, as is the case in AO-CHILDES, but as a set of 5 million sequences with no dependencies across sequences.

Overview of Simulations 2

I am interested in three comparisons: First, I want to show the effect that modification of P_1 on both partitions has on semantic categorization. Specifically, I will compare performance between a group of SRNs trained on equally sized corpora sampled from the same **L**, but where P_1 is set to either 0.5 or 1.0 during corpus generation. In the latter case, no modification is made to **L**, but in the former, only, on average, half of the sequences legal in **L** are actually sampled. If semantic categorization performance is improved when P_1 is 0.5, this would provide strong support to the idea that a reduction in global complexity unrelated to semantic category structure can facilitate acquisition of semantic categories. Secondly, I want to recreate the age-order effect described in chapter 3. To do so, I need to compare the performance of one group of SRNs

trained on artificial input where P_1 increases incrementally, and another group of SRNs trained on the same input, but reversed (P_1 decreases incrementally). This comparison is qualitatively different from the comparison described above, because in each condition, the average complexity is identical, and therefore any difference in performance cannot be due to the average *magnitude* of complexity. The only difference between the two training conditions is the *order* in which complexity varies from start to end of training. My theory predicts that semantic categorization performance will be greater when SRNs are trained on input ordered by increasing complexity. The detailed explanation is provided in chapter 6. Briefly, greater complexity increases the acquisition periods of distinct (singular) dimensions in the input. Consequently, acquisition periods overlap for longer durations, and this increases the risk that information in non-probe word sequences influences semantic differentiation of probe words. Put differently, syntactic constraints semantic differentiation more strongly under conditions of higher complexity. The reason that such constraint is better placed at the start of training is that the SRN primary learns syntactic distinctions during the earliest stages of training, so a constraint imposed by syntactic on semantic differentiation should have a smaller influence on semantic differentiation during this early stage (compared to a later stage of training when more semantic differentiation is occurring).

Third, I want to recreate the conditions exactly as they were when the age-order effect was obtained, except that the input is artificial, and ordered either by increasing or decreasing complexity. Finding a performance improvement in the increasing-complexity condition will be the strongest evidence in support of my theory, because the input is identical, and only the order of training is different between the two conditions.

Simulations 2a: Overall Complexity

I already created three corpora, each with a different number of syntactic categories. These corpora represent the maximum amount of complexity because they were created by sampling from each **L** with P_1=1.0. I derived three additional corpora by sampling from each **L** with P_1=0.5. To understand how this affects **L** associated with each corpus, I plotted each **L** in figure 7.4. The left panels show **L** for the corpora with maximum complexity, while the panels on the right show the same **L** but where half of the +1s in columns corresponding to non-probe words have been replaced with -1s. A legal sequence is shown in red (+1) and an illegal sequence is shown in blue (-1). The meaning of the phrase 'sample from **L** with P_1=0.5' should be now more clear. Essentially, the amount of legal sequences involving non-probe words has been reduced by one half. Because these reduced-complexity corpora are constrained to contain the same number of total sequences as the maximum-complexity corpora, the same sequences re-occur more frequently in the reduced-complexity corpora. An immediate consequence is that co-occurrence frequencies, between a sequence-initial word and a context word are, on average, greater for the reduced-complexity corpora. This drives up the singular values of the term-by-window co-occurrence matrix (term-by-term, in this context) as described in chapter 6.

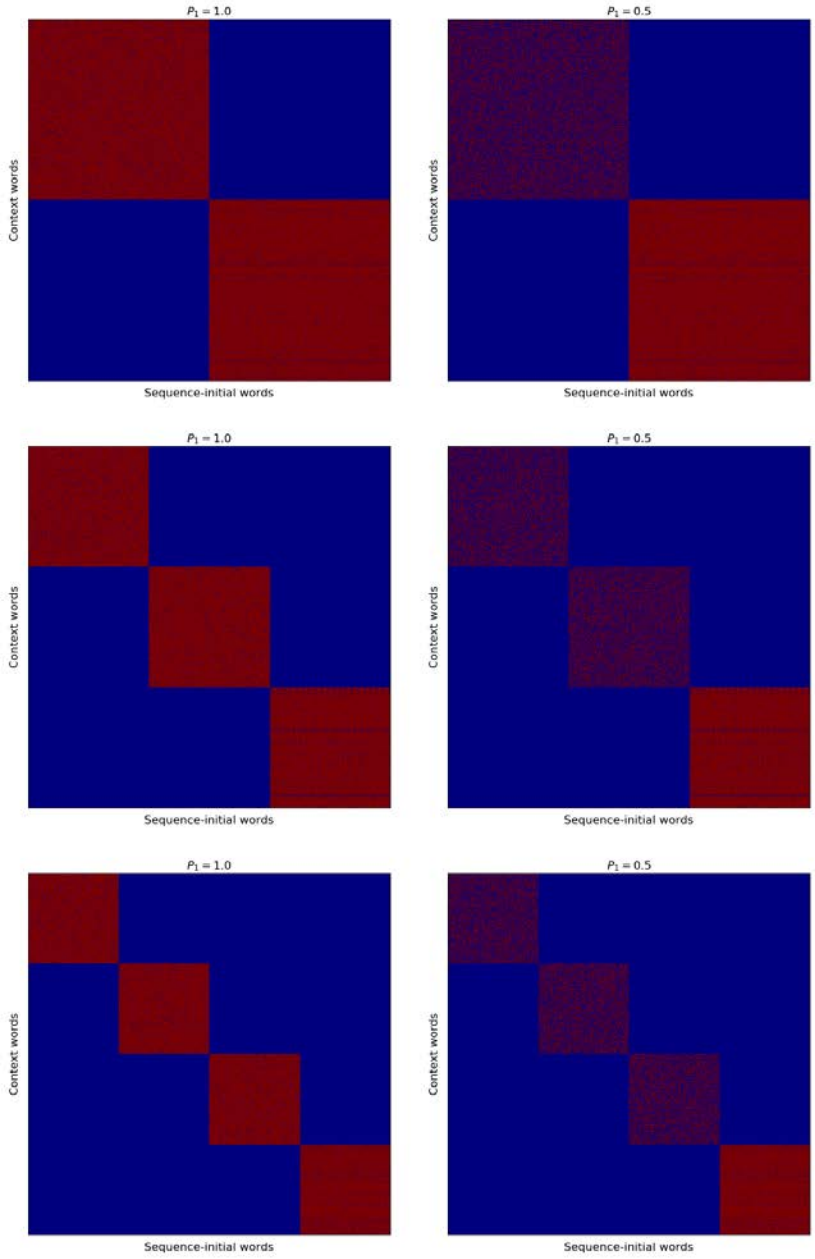

Figure 7.4
The matrix **L** for C_1 (top), C_2 (center), and C_3 (bottom) shown in the left panels. Their reduced-complexity (P_1 reduced from 1.0 to 0.5) counterparts are shown in the right panels. Blue = -1, Red = +1. +1 indicates a legal sequence.

The important difference between each of the corpora pairs, is an overall reduction in complexity, meaning that both partitions in the reduced-complexity corpora are affected. Again, probes are unaffected by the reduction in complexity, so semantic categorization should be

improved when training on the reduced-complexity corpora. The results, shown in figure 7.5, demonstrate that this is the case. While there is no clear performance improvement when there are only 1 or 2 additional syntactic categories (top panels), there is a clear improvement in balanced accuracy when training on the reduced-complexity corpus with 3 added syntactic categories compared to its maximum-complexity counterpart (lower panel). Interestingly, the performance gap in the latter case appears to reduce towards the end of training. It is possible that the gap would close with more training. This would mean that semantic category learning is simply delayed when training on input that is more complex, and does not actually lower asymptotic performance. Eventually, after the SRN has encoded the additional complexity unrelated to probe words, it might still be able to acquire the category structure of probe words at an equal level of performance obtained by the SRN that was not exposed to the full level of complexity. It is very difficult to tease apart whether a model *never* reaches the asymptotic performance achieved by another model, because it is not clear under what circumstances this might be the case, and for how long to continue training.

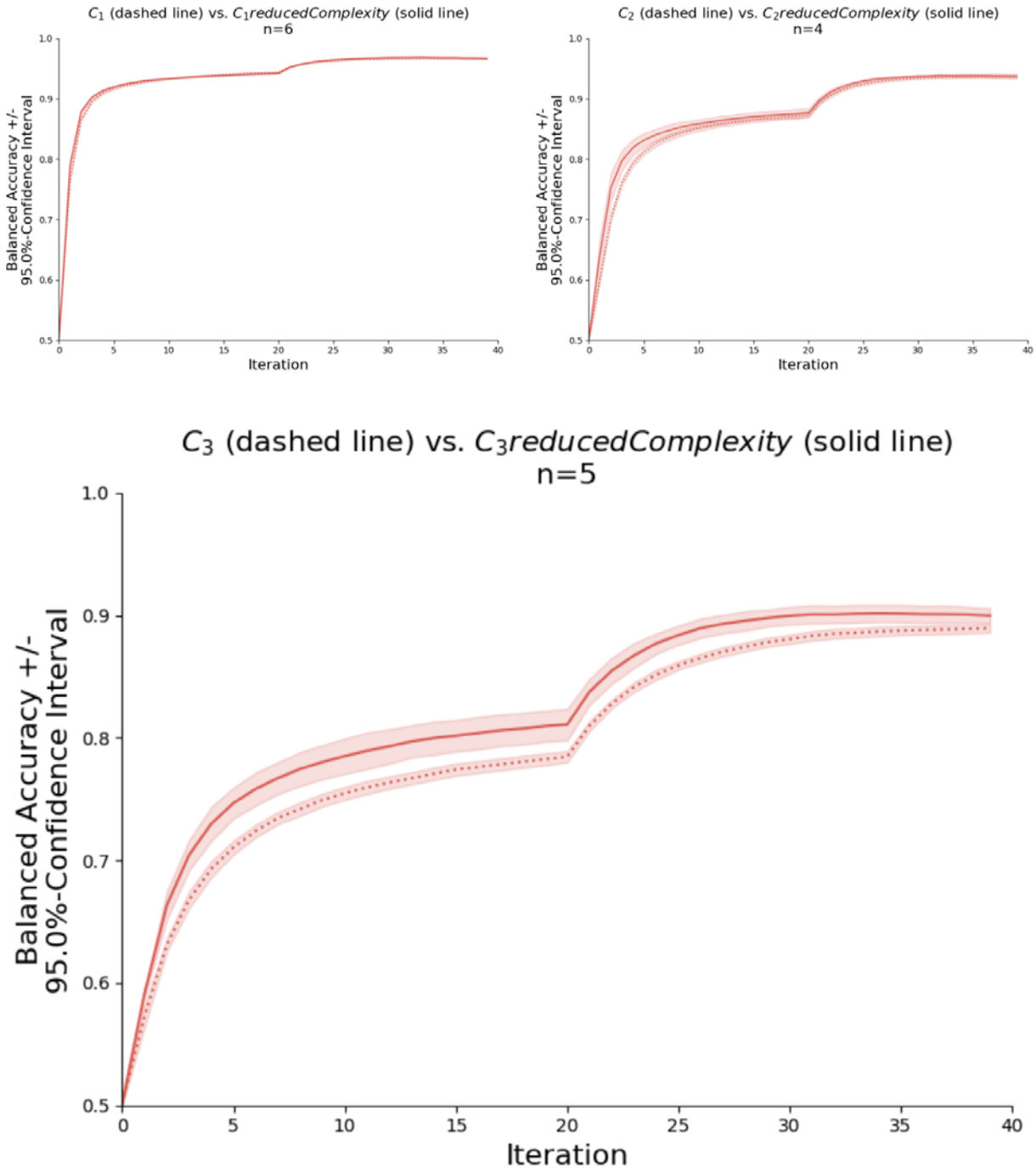

Figure 7.5
Average balanced accuracy as a function of SRN training time (in number of iterations) for SRNs trained on a maximum-complexity corpus (dashed line) and its reduced-complexity counterpart (solid line). Balanced accuracy was computed only for the 5th level in the semantic category hierarchy, in which there are 32 categories.

There are a few additional things to note. First, while I could have plotted balanced accuracy for all 5 levels in the category hierarchy, I did not do so here, because I am most interested in the 5th level in the hierarchy, in which there are 32 categories. This number of categories roughly corresponds to the number of categories (28) used in the simulations in which the age-order effect was observed. Secondly, in all three panels, there is a clear spike in performance at the midpoint. This occurs because both performance has gradually leveled off after iterating 20 times over partition 1, and novel sequences are seen for the first time at the onset of training on partition 2.

Lastly, performance is overall best when the number of syntactic categories is lowest. This is not surprising because a large portion of the representational resources can be devoted to the acquisition of semantic category structure, with little influence exerted by non-probes. But, the reader might wonder why no performance gap was observed when the number of added syntactic categories was lowest (C_1 and C_2). Perhaps the small number of non-probe words in C_1 and C_2 is simply not sufficient to noticeably influence semantic differentiation. It is possible that performance can be improved only in the case in which there is considerable opportunity for non-probe words to influence the course of semantic differentiation of probe words. In this light, the ratio of probe to non-probe words might predict whether a performance gap will occur. Similarly, it is possible that the ratio of the number of *sequences* that are legal within the noun category relative to other syntactic categories determines the magnitude of the performance improvement. Complexity reduction decreases the number of legal sequences that can be formed with non-probe words, while the number for probe words is kept constant. Support for this idea comes from offline simulations in which semantic category structure was generated using a larger mutation probability (increased from 0.01 to 0.03). The effect of this is that there are fewer

legal sequences that can be constructed with probe words (more +1s are 'mutated' to -1s). Applying the same reduction in complexity, which only affects non-probe words, therefore does not have as large an effect on the ratio of legal probe-word sequences to legal non-probe word sequences as before when the mutation rate was lower. Consistent with this idea, I found a smaller performance improvement when training on the reduced-complexity corpus of C_3 compared to training on C_3. Overall, this suggests that what matters most may be the ratio of the complexity of the noun category relative to all other categories. I use complexity to refer, as usual, to the number of legal sequences.

It is important to note that the total number of words in the vocabulary also differs between the three simulations. It is possible that an age-order-like facilitatory effect on performance is (instead) related to the overall number of parameters in the model, or some more complex function of the number of hidden units per vocabulary word. The SRN does not have infinite representational capacity, and therefore the tradeoff between representing syntactic vs. semantic category structure is especially important when the number of parameters is small

Overall, the results of this set of simulations confirm that the routines for both generating the artificial corpora and reducing their complexity do what they are supposed to do.

Simulations 2b: Starting Small

To add incremental structure to the artificial corpora C_1, C_2, and C_3, I simply varied P_{legal} during the random sampling process. To be consistent with the simulations described in chapter 3, in which incremental structure consisted of 2 distinct partitions, I only varied P_{legal} at the point at which half of the 5 million sequences have been sampled. This results in two distinct partitions. To model incremental structure that 'starts small', I left intact partition 2 by setting

P_{legal} to 1.0, but reduced P_{legal} to 0.5 during generation of sequences in partition 1. I did this for C_1, C_2, and C_3, and will refer to the results as $C_{1\_starting\_small}$, $C_{2\_starting\_small}$, and $C_{3\_starting\_small}$.

The resulting corpora are supposed to resemble AO-CHILDES in terms of relative complexity between the first and second half. One can verify this by comparing the number of unique constructions in partitions 1 and 2 of AO-CHILDES to the number of unique sequences in p_1 and p_2 in any of the three corpora. The number of constructions in AO-CHILDES is equivalent to the number of windows, or sequences, that the model sees during training. For this comparison, I picked $C_{2\_starting\_small}$ arbitrarily. AO-CHILDES has 185,087 unique bi-grams in partition 1, and 197,429 unique bi-grams in partition 2, and there are 94,469 bi-grams that are common to both partitions. The numbers are strikingly similar to those computed for $C_{2\_starting\_small}$: There are 178,416 bi-grams in partition 1, and 194,697 bi-grams in partition 2, 89,0966 of which are common to both. Of course there are constructions spanning more than just two words in AO-CHILDES, but $C_{2\_starting\_small}$ contains only 2-word sequences, so comparison is impossible.

There are two more dimensions along which $C_{1\_starting\_small}$, $C_{2\_starting\_small}$, and $C_{3\_starting\_small}$ must be compared with AO-CHILDES. In both cases, a similar pattern should be observed, otherwise the simulations would not be an appropriate test of my theory. The first concerns the singular values associated with the term-by-window co-occurrence matrix computed on each partition. I argued in the previous chapter that a reduction in complexity of the surface structure of any sequential data that can be represented in a term-by-window co-occurrence matrix, must result in singular values whose sum is larger, and verified that this is indeed the case for AO-CHILDES. I will verify that this also holds for the term-by-window co-occurrence matrices computed for all three corpora with incremental structure. If, however, the sum of the singular

values of the first partition of any corpus (less complex than the second partition) is smaller, than this would already invalidate my theory. Previously, I normalized each term-by-window co-occurrence matrix using the L_2 norm before computing their singular-value decomposition, and I will repeat the same procedure here. The results are shown in figure 7.6. For all three corpora, the singular values associated with the first 64 dimensions are larger for the term-by-window co-occurrence matrix computed on the first partition compared to the second partition. Note that while only the first 64 singular values are shown, I verified computationally that the sum of all singular values of the term-by-window co-occurrence matrix is larger when computed on partition 1 compared to partition 2.

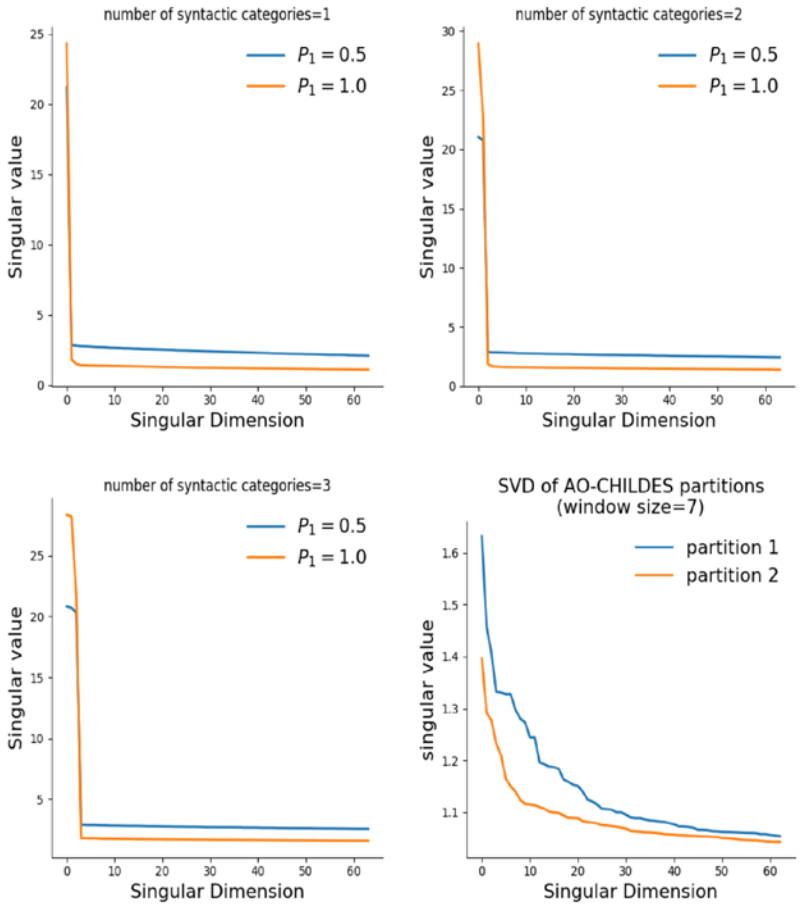

Figure 7.6
Singular values associated with the first 64 singular dimensions obtained via singular value decomposition of term-by-window co-occurrence matrices of $C_{1\_starting\_small}$, $C_{2\_starting\_small}$, and $C_{3\_starting\_small}$ (top left, top right, lower left panel, respectively). The blue line shows singular values for the first partition (sampled with P_1=0.5), the orange line shows singular values for the second partition (sampled with P_1=1.0). For comparison, singular values are shown for partition 1 and 2 of AO-CHILDES (lower right panel).

Secondly, I must verify that the subsampling procedure used to modify partitions 1 of C_1, C_2, and C_3, to create $C_{1\_starting\_small}$, $C_{2\_starting\_small}$, and $C_{3\_starting\_small}$, did not add any additional information about semantic category membership of probe words. If so, this would violate my theory, which asserts that the age-order effect is not due to any incremental change in the quantity or quality of information about semantic category membership. Evidence for this was

obtained in chapter 4, where I showed that the balanced accuracy, computed on probe word representations obtained from the term-by-window co-occurrence matrix of partition 1 in AO-CHILDES is no different than the balanced accuracy obtained in the same way using partition 2. This same pattern must hold for all three artificial corpora. I repeated the same analysis with $C_{2\_starting\_small}$, and plotted the results in figure 7.7. Each panel shows the balanced accuracy as a function of the number of words sampled from a given partition (partition 1 in blue, partition 2 in orange). The left panel shows the results obtained for AO-CHILDES, and the right panel shows the results obtained for $C_{2\_starting\_small}$. In both cases, the balanced accuracy is approximately equal, which indicates that there is no more information about semantic category membership in partition 1 compared to partition 2. Note, that while I have only shown results for $C_{2\_starting\_small}$, the same trend was observed for the other two artificial corpora.

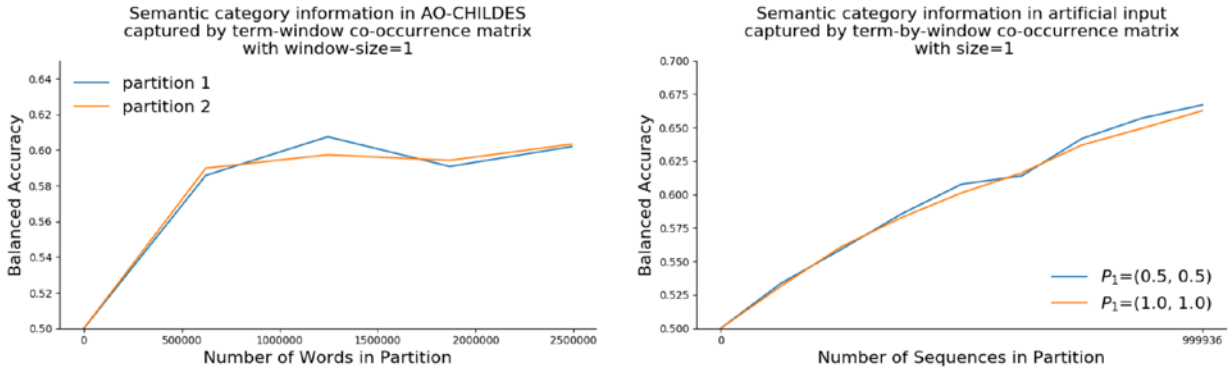

Figure 7.7
Balanced accuracy as a function of the number of words sampled from each partition (partition 1 in blue, partition 2 in orange) in AO-CHILDES (left panel) and in $C_{2\_starting\_small}$ (right panel).

At this point, it might also be revealing to inspect the similarity between probe-words. The probe-word category can be thought of as being similar to nouns in AO-CHILDES. There, we observed, that the average of the pairwise similarities between all nouns decreases from

partition 1 to partition 2. I commented that this in line with my theory, because the gradual reduction in similarity is an indicator that semantic differentiation within the noun category becomes less constrained during age-ordered training. As nouns are becoming distributionally less similar to each other in the input, there is more room for semantic differentiation of probe words in the 'distributional' space of the model. This is because representations of nouns are moving farther apart, providing more space for movement of probe words along semantically distinct trajectories. While low distributional similarity between nouns is useful at later stages during training, high similarity during early stages of training also provides a unique benefit to SRNs trained in age-order. The category NOUN would be acquired faster, by pulling representations of nouns close together in representational space. With probe words clustered closely in representational space, they are in an ideal position to begin semantic differentiation. High and low distributional similarity of nouns therefore both have a role to play in facilitating semantic category learning. However, when their timing is reversed during training, their effect on semantic differentiation is instead counter-productive. Low distributional similarity between nouns during early training does not pull probe words close together in representational space, and therefore does not position them suitably for semantic differentiation. Moreover, high distributional similarity during a later stage would constrain semantic differentiation, rather than help it along. Naturally, I wanted to know whether the artificial 'starting small' corpora I have generated show a pattern of high-to-low distributional similarity between nouns (probe words). To do so, I computed the term-by-window co-occurrence matrix for partition 1 and 2 of each of the three artificial corpora. I used SVD to reduce the dimensionality of the row vectors to either 32, 64, 128, or 256. Next, I obtained representations for all syntactic categories, including category which contains the probe words, and computed the average pairwise cosine similarities

between members of each category. I plotted the results for $C_{3\_starting\_small}$ in figure 7.8. Note, that the pattern shown in the figure holds for all three artificial corpora. In alignment with my predictions, I found that the distributional similarity of nouns (shown in red) decreases from partition 1 to partition 2. Interestingly, this trend only holds when the number of SVD modes retained during dimensionality reduction is kept small (compare upper left to lower right panel). Because the SVD modes explaining the largest amount of variance presumably carry syntactic, rather than semantic information, one interpretation is that the nouns (probe words) in partition 1 are 'syntactically more distinct' from the other syntactic categories than in partition 2. The removal of half of the legal sequences in each of the three syntactic categories in partition 1 reduced the 'integrity' of each of the three syntactic categories relative to the nouns. Remember that the semantic category structure of probe words is identical between partitions, so this pattern of results cannot be due to a semantic effect.

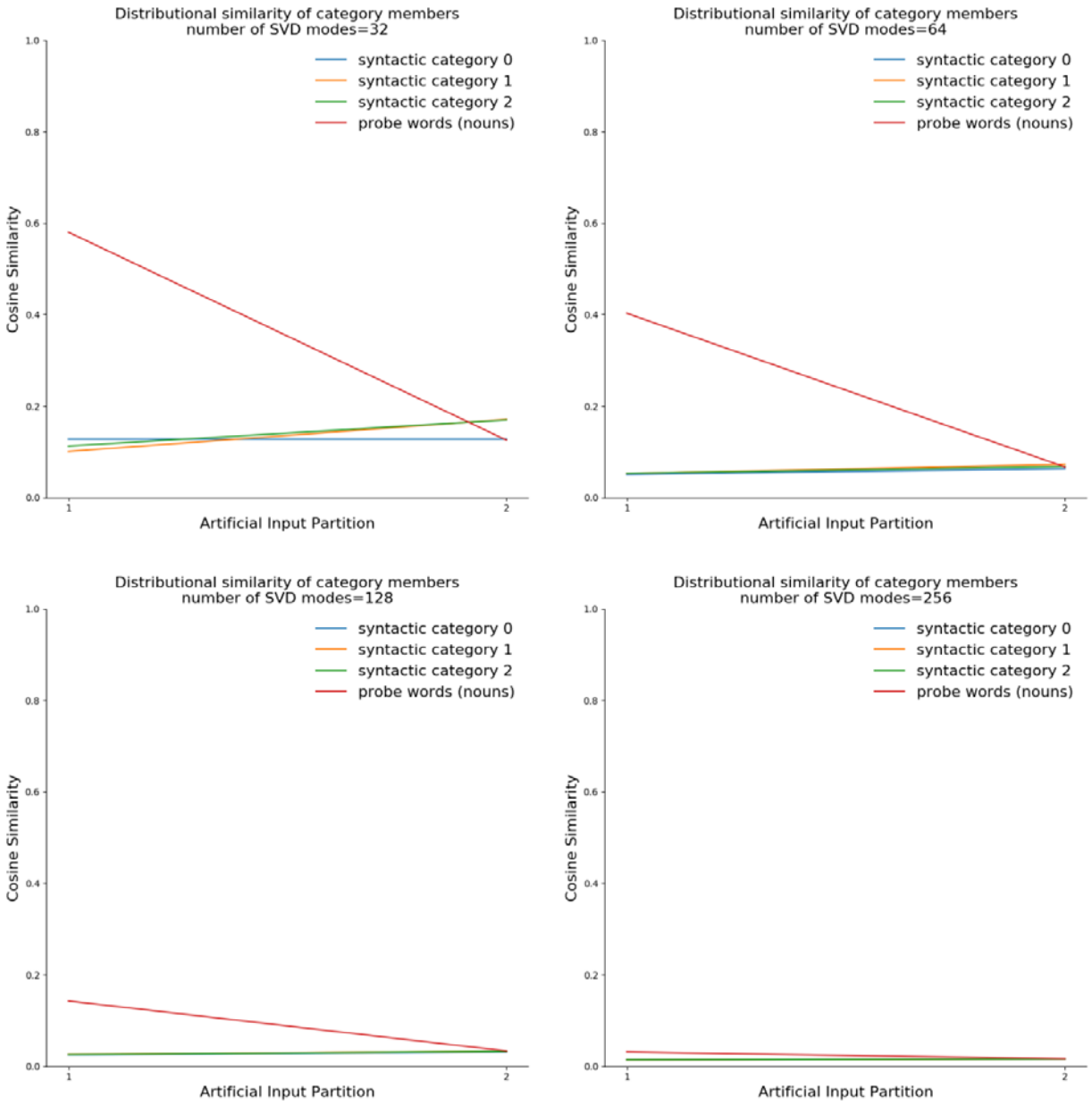

Figure 7.8
Average of pairwise cosine similarities between members of 4 syntactic categories in partition 1 and 2 of $C_{3\_starting\_small}$. Before computing similarities, the dimensionality of the term-by-window co-occurrence matrix was reduced using SVD. The singular dimensions associated with the largest 32, 64, 128, and 256 singular values were retained (top right, top left, lower left, lower right, respectively).

I trained at least eight SRNs on the maximum-complexity corpora and their corresponding ‘starting-small’ versions. I tracked balanced accuracy in consecutive intervals

during training, as described before, and plotted the results in figure 7.9. The main question, to keep in mind with this set of simulations, is whether the performance benefit provided by a reduction in complexity (observed in simulations 2a above) is persistent when limited to partition 1. Put differently, will the performance improvement provided by training on reduced-complexity input during the first half of training last until the end of training? If not, then the results would not align with the age-order effect, which is characterized by an early performance gap that persists until the end of training. Because my theory predicts that the conditions in these simulations is favorable for a persistent performance improvement, failure to find a persistent improvement would weaken the validity of my theory. In the first panel (top left), corresponding to the corpora with only 1 added syntactic category, the two balanced accuracy trajectories are almost identical. Addition of a second syntactic category (top right panel, comparison between C_2 and $C_{2\_starting\_small}$) a performance difference is detectable, and addition of a third syntactic category (lower panel, comparison between C_3 and $C_{3\_starting\_small}$) results in the largest difference in performance. The results are consistent with those obtained for simulations 2a: The latter two cases provide strong evidence that a reduction in complexity unrelated to semantic category structure can noticeably improve semantic category learning. More importantly, the performance improvement observed in the latter two simulations persists until the end of training, despite training on maximum-complexity input last.

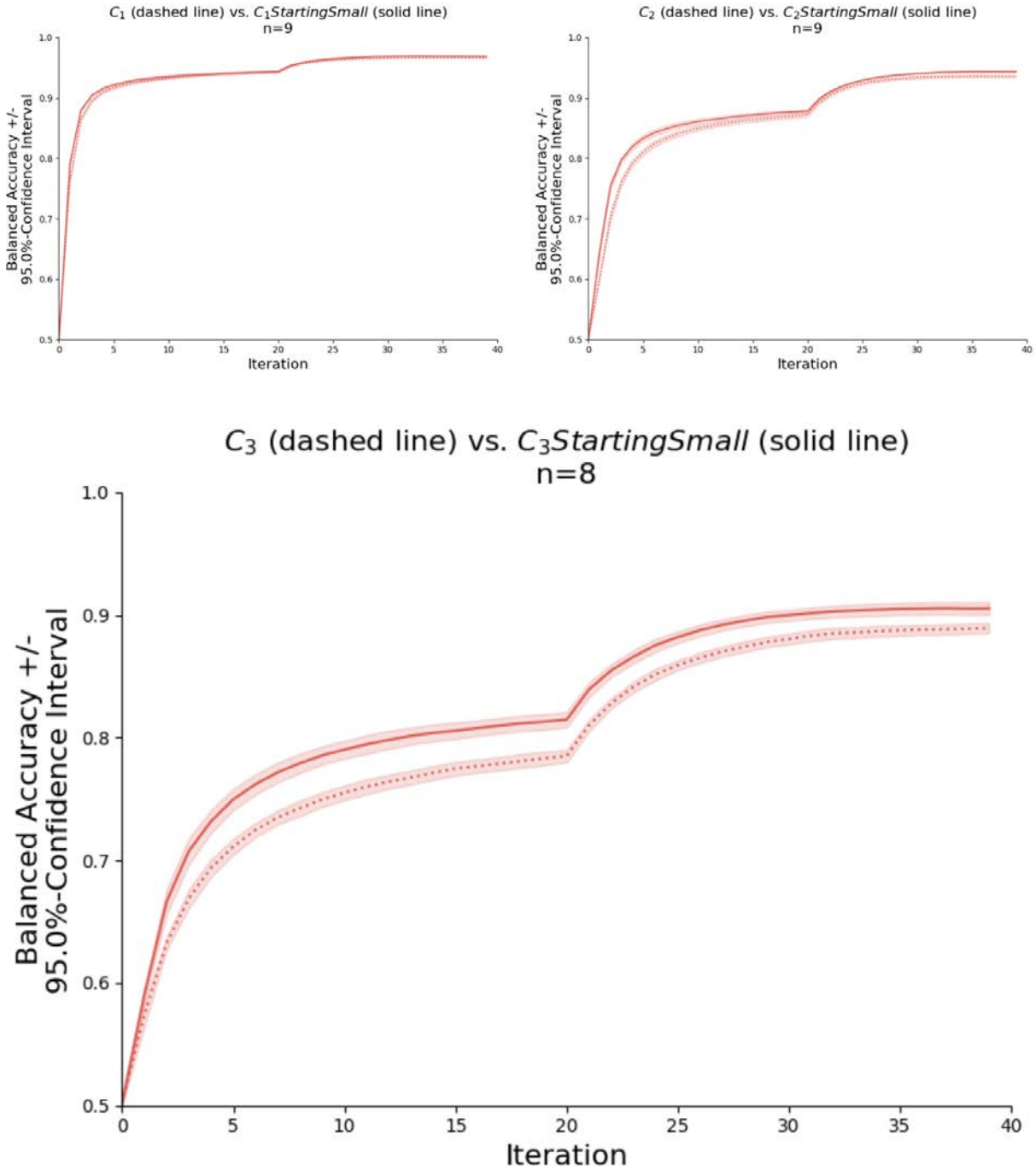

Figure 7.9
Average balanced accuracy as a function of SRN training time (in number of iterations) for SRNs trained on a maximum-complexity corpus (dashed line) and a 'starting-small' corpus (solid line). Balanced accuracy was computed only for the 5th level in the semantic category hierarchy, in which there are 32 categories.

Notice that I did not compare performance of SRNs trained on partition 1 first and partition 2 last to SRNs trained on partition 2 first and partition 1 last. This would be more consistent with the simulations described in chapter 3, in which training on AO-CHILDES occurred either in age-order or reverse age-order. Why did I choose differently? Comparing performance in a 'starting small' condition to performance in a non-'starting small' condition rather than a 'starting large condition is a more direct test of whether the reduction in complexity in partition 1 can improve semantic categorization performance. If, instead, I had chosen the comparison that is more consistent with the one described in chapter 3 (increasing vs. decreasing complexity rather than increasing vs. fixed complexity), I would not be certain that a performance improvement in the 'starting small' (increasing complexity) condition is not in fact due to a decrease in performance in the 'starting large' (decreasing complexity) condition. A reduction in performance in the latter condition may be due to interference that results from acquisition of knowledge in partition 2 which is no longer useful during training on partition 2. Comparing performance of SRNs trained either on input partitions ordered by increasing or decreasing complexity is the subject of the next set of simulations (2c).

The testing conditions above represent those conditions most favorable for an age-order-like performance improvement to occur under my theory. Because I found such an improvement, I have demonstrated that my theory has some predictive power. While these results support the validity of my theory, more work needs to be done to show that it holds in other conditions, with different input and/or different models.

Simulations 2c: Order Matters

In this last set of simulations, I wanted to recreate as best as possible the conditions in which the age-order effect was observed, but where AO-CHILDES has been replaced with an artificial corpus with incremental structure. I chose to use $C_{3\_starting\_small}$ because the number the number of unique words corresponds to the vocabulary size used to train on AO-CHILDES (4,096). It is also the corpus for which the largest 'starting small' effect was observed in the previous section. There, I found that training on reduced-complexity input during the first half of training improves semantic categorization performance both early during training and that this performance improvement persists until the end of training. While this is evidence of the effectiveness of 'starting small', I have not shown that a similar performance improvement cannot be accomplished by 'ending small'. If so, then order does not matter, and complexity reduction, no matter when it occurs in the input (and therefore during training) would be an effective intervention. However, I have argued previously, that order does matter. In fact, my theory is not about whether complexity reduction by itself is beneficial for semantic category learning, but about the order in which complexity is experienced by the model during training. Specifically, it predicts that training on input ordered by increasing complexity improves semantic categorization performance compared to training on input ordered by decreasing complexity. This has to do with the sequential nature of progressive semantic differentiation. Once the model has acquired the most prominent singular dimensions of the input, a reduction in complexity can do very little to reorganize the model's already differentiated representational space.The benefit of a reduction in complexity on semantic category learning is effective only when applied to input that a model first trains on; little or no benefit may be reaped by a model which is already in a more advanced stage of training. In sum, my theory predicts that training on

partition 1 of $C_{3\_starting\_small}$ last will not result in a spike in performance which might eliminate the performance gap observed in simulations 2b. The results of the simulations is shown in figure 7.10. As predicted, the average balanced accuracy for SRNs trained on partitions of $C_{3\_starting\_small}$ ordered by increasing complexity (solid line) is larger than the average balanced accuracy for SNRs trained in the reversed order (dashed line) at all evaluation time points (except at the beginning of training). This is strong evidence in support of the idea that training on input with reduced complexity is facilitatory only when the model has not yet had any prior learning experiences. Early experience with reduced-complexity input appears to scaffold future learning experiences.

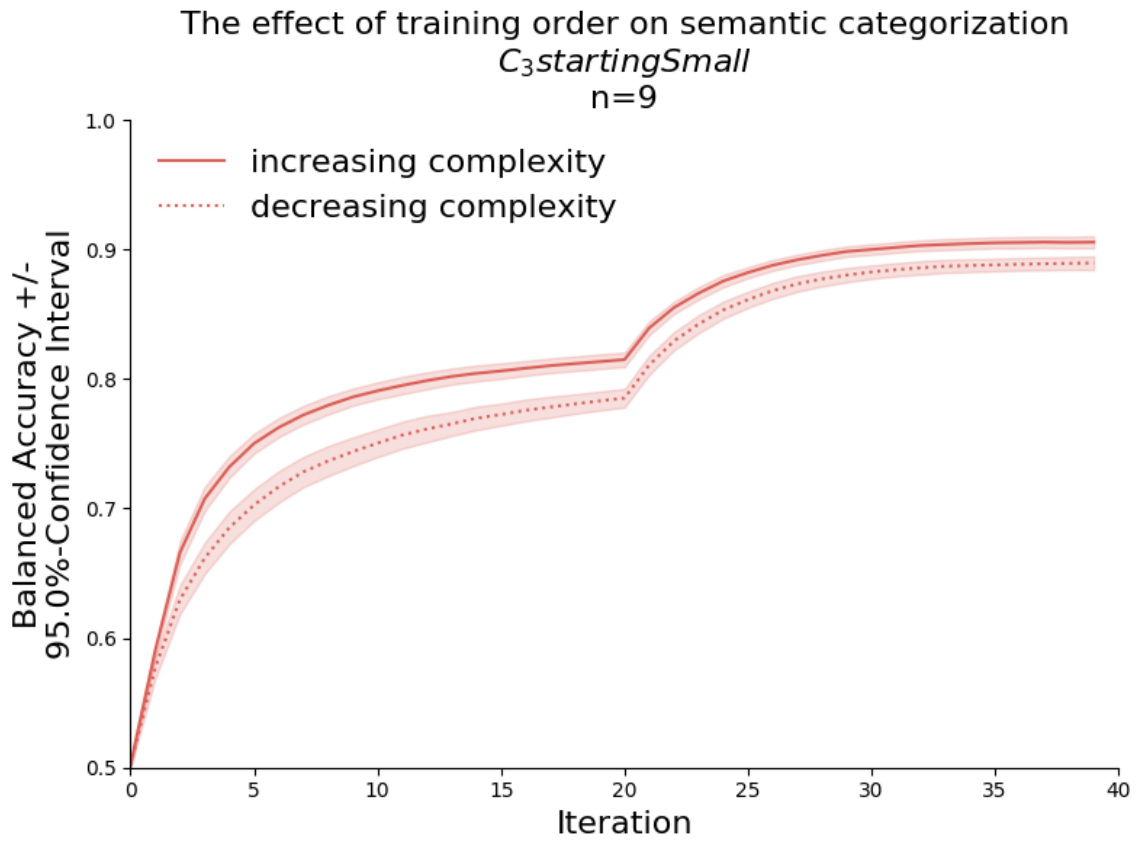

Figure 7.10
Average balanced accuracy as a function of SRN training time (in number of iterations) for SRNs trained on $C_{3\_starting\_small}$ in order of increasing complexity (solid line) and decreasing complexity (dashed line).

The importance of the early training experiences on the SRN's performance found here may not be restricted to the SRN or sequential input like artificial or natural language. It would be interesting to investigate whether similar performance improvements can be obtained in different domains, or tasks. I predict that similar performance improvements should be possible in situations in which the input is composed of a complex subcategory structure where only a subset of lower-level categories are of interest. But the implications do not stop with models implemented on a computer; instead, the input to a human learner is organized incrementally across the lifespan and it would be surprising if the learning system did not take advantage of this fact to speed acquisition of not only semantic category development, but development of motor skills, reasoning, and knowledge learning.

Limitations

The simulations in this chapter were designed to test a theory of the age-order effect. This means that any simulations designed to capture the age-order effect using artificial input, should resemble as closely as possible the original set of simulations. Because the artificial input I have developed is a distant departure from English, it is possible that the results observed in this chapter have no bearing on those observed for AO-CHILDES. I have made several choices when designing the artificial corpora, but have not thoroughly tested their implications. For example, all probe words in the artificial corpora were considered nouns. However, in AO-CHILDES, probe words only represent a small subset of the total number of nouns in the vocabulary. Additionally, AO-CHILDES contains proper nouns, and pronouns which occur in noun-like contexts but are distinct from nouns and probe words. In fact, syntactic category structure in AO-

CHILDES is far more heterogeneous compared to the artificial corpora C_1, C_2, and C_3. In all three, each syntactic category is of the same size and words that are members of a category only occur with context words that are unique to the category. This is most likely not the case in English, where nouns and verbs may share context words (defined here as right adjacent neighbors). Moreover, the density of members of different syntactic categories changes incrementally across AO-CHILDES (e.g. nouns occur more frequently in partition 1, verbs occur more frequently in partition 2). Subtler changes like this in the syntactic category structure are not captured by the artificial corpora.

In the original set of simulations, the age-order effect was observed for SRNs trained with 7 backpropagation-through-time steps (fewer steps reduce the age-order effect), but all sequences in the toy input used in the simulations described in this chapter were of size 2. Because it is possible that the age-order effect might be explained by invoking a systematic difference in the distance of semantic dependencies between the two partitions of AO-CHILDES, the results obtained here may hold little explanatory value. In fact, offline analyses revealed that the age-order effect does not manifest unless backpropagating for more than 4 time steps. Put differently, training SRNs with word sequences 4 words long or less does not result in an age-order effect. It is possible that more complex learning dynamics emerge when backpropagating for more than just 1 time-step, as is done when training on the artificial corpora, and that this makes all the difference. If this is the case, it would be challenging, but essential, to design artificial input where categories are defined in terms of sequential regularities at multiple distances.

CHAPTER 8: RELATIONSHIP TO LEARNING IN INFANTS

The theory developed in this work was designed to explain the behavior of the simple recurrent network, but the ultimate goal of this work is to understand the distributional learning system that is known to guide semantic and syntactic category learning in infants. The goal of this chapter is to relate the theory presented in this work to statistical learning in infants. Concretely, I will discuss two questions: First, what findings in the infant behavioral literature support the theory? Second, is the SRN a good model of infant distributional learning?

Methodological concerns

Although the body of research on infant distributional learning is large, the possibility that sensitivity to sequential regularities in the speech stream actually contributes to semantic category learning has received little prior investigation. Most of the previous work has either focused on speech segmentation or grammatical category learning (Aslin, Saffran & Newport, 1998; Fiser & Aslin, 2002; Kirkham, Slemmer & Johnson, 2002; Pelucchi, Hay & Saffran, 2009; Saffran, Aslin & Newport, 1996). Even in the field of semantic knowledge acquisition more broadly, most researchers study semantic knowledge in terms of perceptual similarity between *referents*, and not in terms of distributional similarity between *labels*. This shortage of studies of how semantic categories are acquired makes it difficult to connect the SRN to infant development. Specifically, no good data exists with which to compare the semantic category judgements of the SRN to that of infants. That is because semantic category judgements in infants would not be tied exclusively to knowledge about linguistic units, but to knowledge about their referents in the real world. The SRN is trained exclusively on linguistic input, and cannot

base its judgements on perceptual similarity or distributional similarity of objects in the real world. This is not a limitation of the SRN, as it can be trained on this kind of information too; rather this is a methodological limitation. To properly compare the SRN's semantic category judgements to those of infants, both would need to be 'trained' on an artificial language with artificial probe words and semantic categories that share no resemblance to entities in the real world. Otherwise, infants would bring to the table a wealth of knowledge from non-linguistic or non-distributional sources (e.g. knowing that lions have four legs, or that a table is solid). If this were done, such a comparison would be limited by its scope. There would be no straightforward way to scale up the artificial language to the level of complexity of a natural language. Even if there were, an infant would be required to learn an artificial language of similar complexity to a natural language that would have no use to the infant other than being able to participate in a behavioral study. Alternatively, rather than trying to control for different sources of information in the infant, the SRN could be trained on those additional sources. The resulting comparison, however, between SRN and infant, would no longer be specifically about how distributional properties of linguistic units are acquired. If such a comparison were to show that SRNs and infants behave similarly, it would be difficult to know why.

A more fruitful way to connect the SRN to infant distributional learning would be to compare syntactic category judgements made by children to those made by an SRN. Because there is no clear equivalent of syntactic categories in the nonlinguistic domain, this would eliminate the possibility that children might take advantage of non linguistic sources of information. While this may be a fruitful avenue of research, the resultant findings may have little bearing on how children acquire semantic categories. There are several reasons why syntactic category judgements should not be treated similarly to semantic category judgements.

First, syntactic category judgements requires precise knowledge of word-order information, whereas word-order is less important for semantic category membership. Second, many have argued that syntactic category knowledge must be constrained and/or supported by a priori syntactic knowledge, whereas knowledge about semantic categories does not. Third, syntactic categories are more straightforwardly formalizable into a system of rules compared to semantic categories. Whether or not the infant distributional learning system treats both kinds of categories similarly will be crucial information for sorting out the role of an SRN-like account of infant distributional learning.

Then, what is the way forward? The best strategy is to compare the SRN to infants in as many ways as possible. There is no single comparison that would be able to tell researchers whether the SRN is a good model of infant distributional learning. or In the meantime, parallels between the SRN and infant distributional learning must be found elsewhere (e.g. in speech segmentation or grammatical classification studies). In the following section, I review a selection of behavioral studies which support the theory of the age-order effect developed in this work. A more general review of studies in support for an SRN-like mechanism of infant distributional learning is beyond the scope of this work. Next, I provide several predictions for testing my theory in behavioral experiments. I will close the chapter by discussing what I take to be the most important limitations of the SRN as a model of infant distributional learning.

Support from behavioral studies

One of the key components of my explanation of the age-order effect in the SRN is that the relatively less complex distributional patterns in partition 1 of AO-CHILDES guide the SRN towards recognizing more complex distributional patterns during training on partition 2. Is there

any evidence that humans benefit from experiencing input with less complex distributional patterns first?

Supporting evidence comes from a statistical learning study, in which English-learning infants were evaluated in their ability to segment an Italian speech stream (Lew-Williams, Pelucchi & Saffran, 2011). It was found that familiarization with a sample speech stream alone was not sufficient for allowing infants to detect word boundaries. Instead, successful detection of word boundaries was detected only when infants were familiarized with the same speech in combination with words heard in isolation. Because isolated words are a frequent occurrence in child-directed speech (Brent & Siskind, 2001; Fernald & Morikawa, 1993), the authors suggested that one word utterances may play a role in preparing infants for future language tasks. Indeed, a benefit of exposure to one-word utterances has been shown on word recognition in sentences (Gout, Christophe & Morgan, 2004; Houston & Jusczyk, 2000; Jusczyk & Aslin, 1995) and later vocabulary development (Brent & Siskind, 2001). The authors explain that words spoken in isolation 'pop out' and therefore provide salient markers in fluent speech for word segmentation. Findings such as these support the idea that reducing the difficulty associated with learning some facts about language can facilitate subsequent learning in more difficult situations. The idea that certain relationships 'pop out' and provide anchors for subsequent learning is similar to the idea that training the SRN on less complex linguistic input first provides a 'good start' for learning semantic distinctions between nouns. In chapter 5 I showed that nouns are distributionally more similar in partition 1, and this may help nouns 'pop out'. Having established a stable category for nouns would facilitate learning finer-grained knowledge about individual nouns.

While the work by Lew-Williams, Pelucchi & Saffran (2011) showed that knowledge about single words can influence subsequent knowledge, what about co-occurrence *relations between words*? After all, the theory developed in this work is not about properties that are tied to individual words (e.g. their sound), but about how words are defined in relation to each other. Is there evidence in the behavioral literature that infants benefit from learning simple co-occurrence statistics first? Work by Lany & Gomez (2008) showed that this is the case. The authors asked whether exposure to adjacent dependencies would facilitate learning of related nonadjacent dependencies. Briefly, the experiment was conducted as follows: Infants 12 month of age were familiarized to an artificial grammar consisting of two-word sequences in which words from either category A or B occur in sequence-initial position, and words from either category X or Y occur in sequence-final position. There were two familiarization conditions: In the control group, words in category A and B did not predict the category of the next word, whereas in the experimental condition, words in category A and B did predict the category of the next word (A was consistently paired with X and B with Y). Following an 8 minute familiarization, infants were habituated to 3-word sequences which followed the same structure as those heard in the experimental condition, except that AX and BY sequences were separated by words in a novel category C (e.g. ACB and BCY). During testing, infants were exposed to sequences which violated the nonadjacent dependency seen during habituation. Successful learning of the nonadjacent dependency during habituation was quantified as a significant increase in mean listening time for the test trials compared to the last two habituation trials. infants learned the nonadjacent dependency. Because 10 month old infants typically fail at learning non adjacent dependencies, it was not surprising that infants in the control familiarization condition did not learn the nonadjacent dependency. However, infants in the

experimental condition who were exposed to *adjacent* dependencies between A and X and B and Y, did learn the *nonadjacent* dependency. The authors concluded that learning nonadjacent dependencies is facilitated by exposure to simpler instances of such structure. This means that infants can generalize from their knowledge of simple structural relationships (e.g. A predicts X) to more complex relationships (e.g. A predicts X after some intervening C). This idea is very similar to the theory developed in this work, which asserts that the reduced complexity of speech to younger children provides a 'good start' for learning semantic dependencies in speech to older children, which is more complex.

More support for the notion that prior experience can influence learning comes from an artificial grammar learning study conducted by Marcus & Fernandes (2007). Infants were exposed to sequences in which words in a particular position are repeated (e.g. the pattern ABB or. ABA), and were subsequently tested on their ability to discriminate pattern-consistent sequences from pattern-inconsistent sequences. The authors found that that successful discrimination was contingent on whether infants were exposed to sequences consisting of speech sounds or sequences consisting of artificial tones. Infants were able to discriminate between novel pattern-consistent and pattern-inconsistent sequences only when sequences consisted of speech sounds. The authors concluded that speech is treated as special by infants, in that it can 'catalyze' learning. An alternative interpretation, also provided by the authors, is that speech sounds are already highly familiar to infants, and that familiarity with elements in the input may give learners access to less salient dependencies between those items. On this view, prior exposure to a particular set of items would allow the learner to uncover the most salient dependencies between those items, and thus free cognitive resources for the detection of less salient dependencies.

The importance of the starting conditions has also been recognized in infants' acquisition of adjectives. For example, Mintz & Gleitman (2002) found that novel adjectives were not readily acquired by 36- and 24-month olds without first being provided rich referential and syntactic information about the meanings of the novel adjectives. Specifically, infants only acquired novel adjectives if they were used to modify nouns referring to specific and familiar objects rather than vague objects (e.g. *thing* or *one*). Mintz & Gleitman (2002) concluded that their findings 'favor an account of lexical acquisition in which layers of information become available incrementally, as a consequence of solving prior parts of the learning problem.'

Underlying the theory developed in this work is the idea that constraints on learning are not independent of the input and can emerge as a consequence of the particular kinds of experience of the learner. A word learning study conducted by Smith & Samuelson (2006) provides strong support. The authors proposed that infants' tendency to extend object labels to similarly shaped objects is due to prior experience with object labels, rather than an built-in bias. By acquiring different labels for objects with different shapes, labels guide attentional resources of the infant towards object shape. According to the authors, this attentional shift towards object properties highlighted by object labels explains why infants extend object labels to similarly shaped objects. It is plausible that training the SRN on speech to younger children first can result in a similar attentional shift toward distributional properties of nouns, and thereby allowing it to acquire a larger number of semantic dependencies between nouns.

More behavioral evidence for the role of the input in guiding categorization comes from a visual category learning study conducted by Horst, Oakes & Madole (2005). The authors were specifically interested in how categorization unfolds over time. In a visual familiarization task, 10 month olds were exposed either to exemplars characterized by a common function or

appearance. When learning exemplars characterized by a common function, infants were initially most sensitive to the common feature, and acquired individual features of exemplars later. On the other hand, when learning exemplars characterized by a common appearance, infants were initially most sensitive to the features that were unique to each exemplar, and only learned the common feature later.

Another critical assumption of the theory developed in this work is that the SRN acquires the most abstract categories in the input first (e.g. syntactic categories), before differentiating between items in subordinate categories (e.g. semantic categories). I have provided evidence that the learning dynamics of the SRN are in fact consistent with this idea. But do infants have the same tendency to acquire more abstract, superordinate categories before learning to differentiate between more concrete, subordinate categories? A study conducted by Younger (1990) found that 10-month olds are more likely to judge a novel prototype as more familiar than exemplars previously seen during a familiarization phase. Judgements made by 13-month olds exposed to the same exemplars showed the opposite pattern: Previously seen exemplars were judged as more familiar than an unseen prototype. the author concluded that 13-month olds are more likely to form an abstract prototypical representation of previously seen exemplars and discard idiosyncratic features of exemplars. Over developmental time, more memory resources come online, and it is thought that this allows greater capacity for remembering item-specific features tied to specific exemplars. This is consistent with the learning dynamics of the SRN which initially encodes dimensions shared by all members of a category, and only then begins to remember additional dimensions that distinguish between members of the category. However, the story is more complicated; for example Tomasello (2003) argues that syntactic categories emerge *after* formation of less abstract categories that are more closely tied to specific items. In

his constructionist account of language acquisition, the building blocks of syntactic categories, such as variable slot schemas (e.g. *the _ walks*), need to be in place before syntactic categories can be acquired.

Predictions

If the theory developed in this work is useful for explaining how infants acquire semantic categories, and how this process might benefit from exposure to less complex speech first, then the theory must withstand future empirical testing. What predictions about infant categorization behavior does the theory make that could be tested?

Infants exposed to input that is both noun-rich and in which nouns are used in similar contexts should show signs of greater knowledge of semantic variation of words within the noun category. The experiment could either be correlational in nature or involve exposure to artificial language in which the distributional properties of nouns can be more easily controlled. But, as shown in the simulations described in chapter 6, the distributional property of nouns need not be modified directly. Rather, it suffices to keep nouns as-is and and reduce the number of unique contexts involving non-nouns. How exactly would this be done? For example, the set size of words that are found adjacent to verbs could be reduced. The same reduction could be applied to adjectives, adverbs, conjunctions, and prepositions, and it might be necessary to modify all of these grammatical categories simultaneously. This should shift representational resources towards nouns and facilitate acquisition of a stable representation of nouns. Of course, the meaning of the word ‘noun’ in an artificial language is defined by the researcher, but this is not a problem. In fact, the theory developed in this work does not rely on any special property of nouns. Any (potentially artificially defined) category, for which semantic category distinctions

exist should benefit from such a manipulation. The primary reason that I studied nouns in this work is that they are more frequently used in child-directed speech. The same benefit of age-ordered training should apply to nouns and verbs, because they, too, can be broken down into smaller semantic classes. For example, a verb can be semantically distinguished by whether the activity it refers to involves transference, motion, or contact. Similarly, an adjective can be semantically distinguished by whether the property it refers to involves color, texture, or size. If these experiments were to turn out in favor of the theory, not only would this highlight the role that syntactic complexity of the input plays in distributional learning, but also that syntactic and semantic categories share a common representational mechanism. It is possible that infants represent syntactic and semantic categories in different representational systems, syntactic complexity therefore would not be able to influence how semantic categories are acquired. However, if infants do represent the two kinds of categories in the same or related spaces, then syntactic complexity must influence acquisition of semantic categories.

The theory also predicts that infants should acquire broader semantic distinctions, such as animate vs. inanimate nouns before acquiring finer-grained distinctions such as nouns that refer to birds vs. nouns that refer to insects. But demonstrating this would not be sufficient support for the theory, because there are several reasons why animate vs. inanimate nouns should be distinguished earlier in development. It is a well known finding in the developmental literature that infants treat objects and agents (e.g. people) differently as early as 2 months of age, and it is thought that they acquire two separate cognitive systems, one that deals with social, and another that deals with non-social cognition (Legerstee, 1992). This may provide an early advantage for recognizing nouns referring to animate vs. inanimate objects. Instead, infants must be evaluated using an artificial language with superordinate semantic categories that do not map onto salient

cognitive dimensions that an infant may have already acquired. However, without such correlated cues (e.g. knowing the meanings of words), learning a semantic subcategory structure, where words can be categorized at multiple levels in a hierarchy, may prove too difficult for infants (or even adults).

Limitations & Future Directions

Having discussed some of the behavioral evidence that supports the ideas developed in this work, there are a number of important limitations that need to be considered. The sections below discuss some of the ways in which the SRN is not ideally suited as a model of infant distributional learning. What constraints or additions are needed to fully capture the rich repertoire of the infant's distributional learning system is the subject of ongoing work.

What is the SRN a model of?

In order to incorporate an SRN-like mechanism into an explanation of infant language acquisition, the precise role of the SRN would need to be sorted out. It is not clear whether the SRN is a model of online language processing or conceptual learning, or consolidation. It is ideally suited as a model of online processing and comprehension, were it not for the multiple number of iterations that the SRN requires to perform at its best. It is more plausible to view iterating 20 times over a day's worth of linguistic input (as was done in experiments in chapter 2) as an offline process that happens during sleep consolidation rather than processes linked to online comprehension, or the phonological loop. It is possible to reduce the chunk size of the input that the SRN iterates over, or not iterate all together, and this would make it more suitable as a model of online comprehension. But semantic categorization would suffer, as there is less

opportunity for integrating over large amounts of input. If the SRN were to occupy the role of an online language processor, would it also be responsible for conceptual learning, or would an additional system need to be added? For example, would an additional system for acquiring and representing semantic categories and relations be required or would they be part of the online comprehension system? While the SRN excels at both syntactic and semantic category learning, representing the two kinds of categories in the same representational space is suboptimal. There is considerable support in linguistics and psycholinguistics that syntax and semantics are separate systems (Jackendoff, 2003). Moreover, how would the semantic representations be accessed during production? Would the same SRN be used for comprehension and production? This is not far fetched; Chang, Dell & Bock (2002) developed the 'dual-path' model, at the heart of which the same SRN is involved in both comprehension and production.

It certainly seems that the SRN can potentially play the role of any part of the human language acquisition system. Several extensions and augmentations have been proposed to deal with the unique requirements of various language systems. But how much is too much? While such augmentations can adapt the SRN to any number of situations, such modifications also render it less suitable to others. To move forward, a consensus needs to be reached as to what exactly the role of the SRN is in human language acquisition.

Computations

While it is known that infants as early as 10 months are able to track transitional probabilities between words and syllables, it is not clear what other kinds of co-occurrence statistics children can learn and use. Co-occurrence patterns vary from simple bi-gram statistics to complex non-adjacent dependencies involving higher order n-gram statistics, and it is

unknown at what age or whether children are capable of processing the more complex cues. The SRN can in principle learn any dependency in the input provided they are not too distant. It is possible that the SRN is too expressive and not constrained in ways that children might be.

Moreover, it is unknown how children compute distributional similarity, which is required for categorization of learned items and generalization to novel items. Some work in this area comes from studies of adult grammatical category learning (Reeder, Newport & Aslin, 2013). The authors explored several variables that adult learners might be sensitive to when computing whether a word should be a member of the same grammatical category. To study generalization, the authors measured the difference in participants' ratings of familiar versus novel grammatical strings. It was found that adults are sensitive to the overlap among contexts across words, a systematic gap in overlap of contexts, and the overall number of items they were exposed to. The authors concluded that adults use distributional properties in a principled way when determining whether to generalize.

Biological Plausibility

The theory developed in this work describes how a neural network behaves in response to modifications of the order in which training examples are presented. The theory is thus specific to the learning algorithm used to train neural networks. It may generalize to infants, but only if the same or similar learning algorithm drives distributional learning in infants. An important requirement, thus, is that the learning algorithm must be implementable in neuronal hardware. The algorithm that is most commonly used to train neural networks, and used to train the SRN in this work, is backpropagation. The neurobiological plausibility of backpropagation has often been questioned (Zipser & Andersen, 1988; Crick, 1989; Stork, 1989). However, Xie and Seung

(2003) proved that—under certain conditions—backpropagation is mathematically equivalent to contrastive Hebbian learning, a process that many researchers believe can be implemented in neuronal hardware.

Convergent Cues

The SRN readily learns the distributional patterns in the input, but researchers typically do not find evidence of distributional learning in infants without having provided redundant sources of information ('convergent cues') about category membership. In fact, only a small number of studies so far have found evidence for distributional learning of grammatical categories in the absence of convergent cues, like word meaning or phonology (Mintz, 2002). In other words, children do not appear to use distributional information as their sole source of information when acquiring grammatical categories. Whether this is because it is not sufficient to induce adult-like grammatical category structure is an outstanding question; some have argued this the case (Pinker, 1987). Learners can easily overcome this weakness when additional, redundant sources of information about category membership are available. For example, Lany & Gómez (2008) showed that when multiple cues are available and provide convergent evidence, infants have no trouble using distributional information as a basis for acquiring grammatical categories. Specifically, 12 month-olds dishabituated (in a visual fixation task) to sequences that violated the category structure they were previously exposed to, but only when the familiarization included phonological similarity as a cue to category structure. The kind of phonological cue used was syllable length: In the experimental condition, the category of the first word of each string covaried consistently with the syllable length of the subsequent word, but in the control condition this variation was not kept consistent. Because only children who

were in the experimental condition showed dishabituation to illegal category structures during testing, the authors concluded that distributional analysis does not happen in a vacuum, and requires evidence from other sources to be used for category learning.

How exactly do infants combine linguistic co-occurrence patterns with cues from other domains? For example, prosodic patterns, stress patterns, phonotactic patterns, the referential context of the interaction, and the speaker's eye gaze could all factor into the computation of the infant's distributional learning system. Additional sources of information may highlight which of the sequential regularities identified via distributional analysis are most likely to be psychologically useful cues to category membership. If the SRN were extended to capture information from multiple sources, this would raise many questions about implementation.

Linguistic vs. Non-linguistic Knowledge

The theory described in this work was developed to shed light on the distributional learning mechanism of SRNs trained exclusively on linguistic input. However, the SRN has no access to or knowledge of grounded, embodied, world knowledge that most (but not all) children receive from vision, hearing, touch, taste, and smell. Under more realistic conditions, in which world knowledge is available, the theory may not hold, and therefore would fail to have any bearing on the distributional learning system in infants. World knowledge may interact and modify representations of words as children learn to link words to their meanings. Training the SRN or a different neural network on nonlinguistic sources of information is possible, and there are no a priori reason why co-occurrence statistics of entities in the real world should be treated differently by such a network. Though, in such a network, syntactic complexity may no longer

play as important a role as it does when only linguistic input is available. For infants, speech is probably not the only input that is experienced in a developmentally staged (age-ordered) manner. For example, visual and tactile input is initially constrained to a small set of toys and people (e.g. family members but very few strangers). The number of environments that a young infant experiences is probably smaller and the kinds of environments less varied than those experienced by older infants. A theory of infant distributional learning must take into consideration how changes (e.g. in complexity) in nonlinguistic input channels influence the representations of words.

Symbolic Variables

Some researchers think that distributional learning is not sufficient for explaining how infants acquire the complex compositional structure of a language. For example, Marcus et al. (1999) showed that infants are able to distinguish between two repetition grammars when test items had never before been encountered. Under a distributional learning account, discrimination should have been at chance; therefore, the authors claimed that only an account based on rule-learning can explain this finding. The task involved a brief familiarization phase in which 7-month old infants were exposed to 3-syllable sequences that either followed an ABA or ABB pattern (e.g., *wo fe wo* or *wo fe fe*). In a subsequent test phase, infants were exposed to pattern-consistent or pattern in-consistent sequences consisting of syllables that never occurred during training. Based on differential looking times to consistent vs. inconsistent test sequences, the authors concluded that infants were able to discriminate between the two different repetition grammars, and that they were able to do so because they acquired an abstract rule that operates independently of the items seen during the exposure phase. To succeed at this task, the authors

claimed that it is not sufficient to learn the dependencies between specific training items; rather, an abstract rule based on identity must be acquired, which operates over sequence *positions* rather than *specific items*. Moreover, Marcus et al. (1999) showed that several popular neural networks fail to generalize to novel test sequences, and argued that the reason they failed on this task is because they are fundamentally limited as models of infant language acquisition. The authors argued that neural networks cannot represent symbolic variables, which the authors consider to be necessary for language.

If true, any theory built on observations of neural networks would not be appropriate for explaining how infants acquire language. This is an obstacle to my theory which is based on studies of a neural network. Since the publication of the Marcus et al. (1999) paper, little progress has been made to show that neural networks can in fact represent symbolic variables or learn abstract rules (Alhama & Zuidema, 2019; but see Chang 2002). However, a lot of subsequent work showed that neural networks need not learn abstract, algebraic rules to succeed at the ABB vs. ABA discrimination task. Instead, researchers argued that the neural networks that failed to generalize lacked prior knowledge that repetition is an important aspect of the environment. When the SRN was pre-trained on input consisting of sequences in which some items were systematically repeated before being exposed to the set of sequences used by Marcus et al. (1999), the SRN was able to discriminate between the repetition grammars (Calvo & Colunga, 2003; Seidenberg & Elman, 1999). Thus, it appears that the infant data collected by Marcus et al. (1999) can be accounted for by a distributional learning account, one that does not require symbolic variables or abstract algebraic rules.

Conclusion

What makes the age-order effect a fascinating topic of inquiry is that the SRN - a neural network - benefits from training on input in the order that *children* actually experience it. This link between the SRN and the incrementally structured language environment in which infants are raised suggests that the distributional learning system underlying the infant and the SRN may be governed by similar principles. Moreover, the age-order effect is a reminder of the important role that early experiences have on subsequent learning experiences. While this is well known in the developmental literature, it is not often discussed by neural network researchers, who train their models on data in randomized order. This may be due, in part, to the fact that relatively few neural network studies have documented effects of training order on performance. However, as my corpus analyses revealed, infants do *not* experience the world in random order, and researchers using neural networks to model language acquisition should take note.

While many effects of early experiences on later learning have been documented in infants, there is no evidence that *semantic categorization* benefits from experiencing less syntactically complex speech first. I demonstrated that this is the case in the SRN, and argued that to the extent to which the SRN is a good model of infant distributional learning, infants, too, should benefit in a similar fashion. If true, the theory developed in this work would help to explain why. While the theory is valuable by itself, in providing an explanation of the age-order effect in the SRN, it would be more valuable if it were also to also apply to infants. I discussed several behavioral studies in which crucial aspects of the theory (e.g. that infants acquire complex relationships after exposure to related simpler structures) aligned with behavior exhibited by infants. These demonstrations lend support to the notion that 1) the theory developed to explain the behavior of the SRN may also explain aspects of infant cognition, and

2) that the SRN is a good model of the infant distributional learning system. But there is much more work to be done, as evident in my discussion of the several ways in which the SRN may not be a good fit after all. If the SRN turns out to be insufficient to explain infant distributional learning, then the theory developed in this work would be less likely to apply to infants. While the SRN has received criticism from many language researchers, I think that a ban of the SRN would be premature, and unwarranted. It is probably true that the SRN, *as implemented and trained in this work*, does not capture the richness and complexity of the infant's distributional learning system, but this does not mean that the SRN is inadequate *in principle*. Since the SRN was first introduced, a great number of modifications and augmentations have been introduced to deal with the many criticisms it has received from all fronts in the language research community. For example, the SRN may be augmented with more powerful hidden units that allow it to learn longer distance dependencies (Hochreiter & Schmidhuber (1997), combined with a second SRN to learn mappings between sequential input to sequential output (Sutskever et al., 2014), trained simultaneously on input from multiple domains, and integrated with additional components to enable acquisition of symbolic variables (Chang, 2002).

Whether the theory developed in this work actually applies to infants is a complicated question in part because it is not known whether infant semantic category acquisition actually benefits from simplified speech input. To demonstrate that it does would require raising children who are provided only linguistic input, and are in every other way cut off from experiencing the world. Clearly, this would be unfeasible. Fortunately, there are ways around this issue, such as using artificial language experiments with nonsense words that do not map onto real world concepts. To encourage further inquiry into the ecological validity of the theory, I have provided several predictions for future empirical testing.

REFERENCES

Alhama, R.G. & Zuidema, W. (2019). A review of computational models of basic rule learning: The neural-symbolic debate and beyond. Psychon Bull Rev. doi.org/10.3758/s13423-019-01602-z

Baroni, M., & Lenci, A. (2011, July). How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics* (pp. 1-10). Association for Computational Linguistics.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. J. Mach. Learn. Res. 3, 1137–1155.

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009, June). Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning* (pp. 41-48). ACM.

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*(2), B33-B44.

Broen, P. A. (1972). The verbal environment of the English-learning child. *ASHA Monographs*, *17*.

Buchnik, E., Cohen, E., Hassidim, A., & Matias, Y. (2019). Self-similar Epochs: Value in arrangement. ICML.

Bullinaria, J. A., and Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: a computational study. *Behav. Res. Methods* 39, 510–526. doi: 10.3758/BF03193020

Burgess, C., and Lund, K. (1998). "Modeling cerebral asymmetries in high-dimensional semantic space," in *Modeling Cerebral Asymmetries in High-Dimensional Semantic Space*, eds M. Beeman and C. Chiarello (Hillsdale, NJ: Lawrence Erlbaum Associates), 215–244.

Calvo, F., & Colunga, E. (2003). The statistical brain: Reply to Marcus' The algebraic mind. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 25, No. 25).

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science*, *27*(6), 843-873.

Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. *Cognitive science*, *26*(5), 609-651.

Chen, D., and Manning, C. (2014). "A fast and accurate dependency parser using neural networks," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Doha: Association for Computational Linguistics), 740–750. doi: 10.3115/v1/D14-1082

Clark, A. (1993). *Associative engines: Connectionism, concepts, and representational change*. MIT Press.

Cleeremans, A., and McClelland, J. L. (1991). Learning the structure of event sequences. *J. Exp. Psychol. Gen.* 120, 235–253. doi: 10.1037/0096-3445.120.3.235

Crick, F. (1989). The recent excitement about neural networks. *Nature*, *337*(6203), 129-132.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* 12, 2121–2159.

Eisler, Zoltán, Imre Bartos, and János Kertész. 2007. Fluctuation scaling in complex systems: Taylor's law and beyond. Advances in Physics, pages 89–142

Elman, J. L. (1990). Finding structure in time. *Cogn. Sci.* 14, 179–211. doi: 10.1207/s15516709cog1402_1
Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Mach. Learn.* 7, 195–225. doi: 10.1007/BF00114844
Fernald, A., & Morikawa, H. (1993). Common themes and cultural variations in Japanese and American mothers' speech to infants. *Child development*, *64*(3), 637-656.
Fernald, A., Perfors, A., & Marchman, V. A. (2006). Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. *Developmental psychology*, *42*(1), 98.
Fernald, A., & Kuhl, P. (1987). Acoustic determinants of infant preference for motherese speech. *Infant behavior and development*, *10*(3), 279-293.
Firth, J. R. (1957). *Studies in Linguistic Analysis*. Oxford: Blackwell.
Fisher, C., Gertner, Y., Scott, R. M., and Yuan, S. (2010). Syntactic bootstrapping. *Wiley Interdiscip. Rev. Cogn. Sci.* 1, 143–149. doi: 10.1002/wcs.17
Fodor, J. A., and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition* 28, 3–71. doi: 10.1016/0010-0277(88)90031-5
Foushee, R., Griffiths, T., & Srinivasan, M. (2016). Lexical Complexity of Child-Directed and Overheard Speech: Implications for Learning. In *CogSci*.
Furrow, D., Nelson, K., & Benedict, H. (1979). Mothers' speech to children and syntactic development: Some simple relationships. *Journal of child language*, *6*(3), 423-442.
Gleitman, L. R., Newport, E. L., & Gleitman, H. (1984). The current status of the motherese hypothesis. *Journal of child language*, *11*(1), 43-79.
Graf Estes K, Evans JL, Alibali MW, Saffran JR. Can infants map meaning to newly segmented words? Statistical segmentation and word learning. Psychol Sci. 2007;18:254–260
Graves, A., Bellemare, M. G., Menick, J., Munos, R., & Kavukcuoglu, K. (2017, August). Automated curriculum learning for neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (pp. 1311-1320). JMLR. org.
Gallaway, C., & Richards, B. J. (Eds.). (1994). *Input and interaction in language acquisition*. Cambridge University Press.
Gershman, S. J., and Tenenbaum, J. B. (2015). "Phrase similarity in humans and machines," in *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (Austin, TX: Cognitive Science Society).
Gleitman, L. (1990). The structural sources of verb meanings. *Lang. Acquis.* 1, 3–55. doi: 10.1207/s15327817la0101_2
Golinkoff, R. M., & Alioto, A. (1995). Infant-directed speech facilitates lexical learning in adults hearing Chinese: Implications for language acquisition. *Journal of Child Language*, *22*(3), 703-726.
Gout, A., Christophe, A., & Morgan, J. L. (2004). Phonological phrase boundaries constrain lexical access II. Infant data. *Journal of Memory and Language*, *51*(4), 548-567.
Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing.
Hart, B., and Risley, T. R. (2003). The early catastrophe: the 30 million word gap by age 3. *Am. Educ.* 27, 4–9.
Harris, Z. S. (1954). Distributional structure. *Word* 10, 146–162. doi: 10.1080/00437956.1954.11659520

Heafield, K., Pouzyrevsky, I., Clark, J. H., & Koehn, P. (2013). Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Vol. 2, pp. 690-696).

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Horst, J. S., Oakes, L. M., & Madole, K. L. (2005). What does it look like and what can it do? Category structure influences how infants categorize. *Child Development*, *76*(3), 614-631.

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(5), 1570.

Huebner, Philip A., and Jon A. Willits. "Structured semantic knowledge can emerge automatically from predicting word sequences in child-directed speech." *Frontiers in Psychology* 9 (2018): 133.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental psychology*, *27*(2), 236.

Jackendoff, R. (2003). Précis of foundations of language: brain, meaning, grammar, evolution. *Behavioral and Brain Sciences*, *26*(6), 651-665.

Jones, M. N., Kintsch, W., and Mewhort, D. J. (2006). High-dimensional semantic space accounts of priming. *J. Mem. Lang.* 55, 534–552. doi: 10.1016/j.jml.2006.07.003

Jones, M. N., and Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychol. Rev.* 114, 1–37. doi: 10.1037/0033-295X.114.1.1

Jusczyk, P. W., & Aslin, R. N. (1995). Infants′ detection of the sound patterns of words in fluent speech. *Cognitive psychology*, *29*(1), 1-23.

Kahneman, D., and Tversky, A. (1972). Subjective probability: a judgment of representativeness. *Cogn. Psychol.* 3, 430–454. doi: 10.1016/0010-0285(72)90016-3

Kirchhoff, K., & Schimmel, S. (2005). Statistical properties of infant-directed versus adult-directed speech: Insights from speech recognition. *The Journal of the Acoustical Society of America*, *117*(4), 2238-2246.

Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* 104, 211–240. doi: 10.1037/0033-295X.104.2.211

Lany, J., & Gómez, R. L. (2008). Twelve-month-old infants benefit from prior experience in statistical learning. *Psychological Science*, *19*(12), 1247-1252.

Lany, J., and Saffran, J. R. (2010). From statistics to meaning: Infants' acquisition of lexical categories. *Psychol. Sci.* 21, 284–291. doi: 10.1177/0956797609358570

Legerstee, M. (1992). A review of the animate-inanimate distinction in infancy: Implications for models of social and cognitive knowing. *Early Development and Parenting*, *1*(2), 59-67.

Lew-Williams, C., Pelucchi, B., & Saffran, J. R. (2011). Isolated words enhance statistical language learning in infancy. *Developmental Science*, *14*(6), 1323-1329.

Lieven, E. V. (1994). Crosslinguistic and crosscultural aspects of language addressed to children.

Lund, K., and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. Behav. Res. Methods Instrum. Comput. 28, 203–208. doi: 10.3758/BF03204766

MacWhinney, B. (2000). The Childes Project: Tools for Analyzing Talk, 3rd Edn. Mahwah, NJ: Lawrence Erlbaum Associates.
Marcus, G. F. (1998). Rethinking eliminative connectionism. Cogn. Psychol. 37, 243–282. doi: 10.1006/cogp.1998.0694
Marcus, G. F., Fernandes, K. J., & Johnson, S. P. (2007). Infant rule learning facilitated by speech. *Psychological science*, *18*(5), 387-391.
Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77-80.
Mikolov, T. (2012). Statistical language models based on neural networks. *Presentation at Google, Mountain View, 2nd April*, *80*.
Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. arXiv:1301.3781
Mikolov, T., Joulin, A., Chopra, S., Mathieu, M., and Ranzato, M. A. (2014). Learning longer memory in recurrent neural networks. arXiv:1412.7753
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). "Distributed representations of words and phrases and their compositionality," in Proceedings of the 26th International Conference on Advances in Neural Information Processing Systems, Lake Tahoe, NV, 3111–3119.
Mintz, T. H., & Gleitman, L. R. (2002). Adjectives really do modify nouns: The incremental and restricted nature of early adjective acquisition. *Cognition*, *84*(3), 267-293.
Nelson, D. G. K., Hirsh-Pasek, K., Jusczyk, P. W., & Cassidy, K. W. (1989). How the prosodic cues in motherese might assist language learning. *Journal of child Language*, *16*(1), 55-68.
Newport, E., Gleitman, H., & Gleitman, L. (1977). Mother, Id rather do it myself: Some effects and non-effects of maternal speech style.
Olney, A. M., Dale, R., and D'Mello, S. K. (2012). The world within Wikipedia: an ecology of mind. *Information* 3, 229–255. doi: 10.3390/info3020229
Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: global vectors for word representation. *EMNLP* 14, 1532–1543. doi: 10.3115/v1/D14-1162
Pereira, F., Gershman, S., Ritter, S., and Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cogn. Neuropsychol.* 33, 175–190. doi: 10.1080/02643294.2016.1176907
Pine, J. M. (1994). The language of primary caregivers.
Pinker, S. (1987). The bootstrapping problem in language acquisition. *Mechanisms of language acquisition*, 399-441.
Pinker, S., and Prince, A. (1988). On language and connectionism: analysis of a parallel distributed processing model of language acquisition. *Cognition* 28, 73–193. doi: 10.1016/0010-0277(88)90032-7
Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2014). Look who's talking: speech style and social context in language input to infants are linked to concurrent and future speech development. *Developmental science*, *17*(6), 880-891.
Rafferty, A.N., & Griffiths, T.L. (2010). Optimal Language Learning: The Importance of Starting Representative.
Reeder, P. A., Newport, E. L., & Aslin, R. N. (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. Cognitive psychology, 66, 30-54.

Richards, B. J. (1994). Child-directed speech and influences on language acquisition: Methodology and interpretation.

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., et al. (2004). Structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychol. Rev.* 111, 205–235. doi: 10.1037/0033-295X.111.1.205

Rogers, T. T., and McClelland, J. L. (2008). Précis of semantic cognition: a parallel distributed processing approach. *Behav. Brain Sci.* 31, 689–714. doi: 10.1017/S0140525X0800589X

Rohde, D. L., & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small?. *Cognition*, *72*(1), 67-109.

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science* 274, 1926–1928. doi: 10.1126/science.274.5294.1926

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, *116*(23), 11537-11546.

Schieffelin, B. B., & Ochs, E. (Eds.). (1986). *Language socialization across cultures* (No. 3). Cambridge University Press.

Seidenberg, M.S. and Elman, J.L. (1999). Generalization, rules, and neural networks: A simulation of Marcus et al. http://crl.ucsd.edu/~elman/Papers/MVRVsimulation.html

Singh, L., Nestor, S., Parikh, C., & Yull, A. (2009). Influences of infant-directed speech on early word recognition. *Infancy*, *14*(6), 654-666.

Smith, L. B., & Samuelson, L. (2006). An attentional learning account of the shape bias: Reply to Cimpian and Markman (2005) and Booth, Waxman, and Huang (2005).

Smith, E. E., Shoben, E. J., and Rips, L. J. (1974). Structure and process in semantic memory: a featural model for semantic decisions. *Psychol. Rev.* 81, 214–241. doi: 10.1037/h0036351

Snow, C. E. (1972). Mothers' speech to children learning language. *Child development*, 549-565.

Snow, C. E., & Ferguson, C. A. (1977). Talking to children.

Steyvers, M., and Tenenbaum, J. B. (2005). The Large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cogn. Sci.* 29, 41–78. doi: 10.1207/s15516709cog2901_3

Stork, D. G. (1989, June). Is backpropagation biologically plausible. In *International Joint Conference on Neural Networks* (Vol. 2, pp. 241-246). Washington, DC: IEEE.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (pp. 3104-3112).

Syrett, K., and Lidz, J. (2010). 30-month-olds use the distribution and meaning of adverbs to interpret novel adjectives. *Lang. Learn. Dev.* 6, 258–282. doi: 10.1080/15475440903507905

Tanaka-Ishii, K., & Kobayashi, T. (2018). Taylor's law for linguistic sequences and random walk models. *Journal of Physics Communications*, *2*(11), 115024.

Tardif, T., Shatz, M., & Naigles, L. (1997). Caregiver speech and children's use of nouns versus verbs: A comparison of English, Italian, and Mandarin. *Journal of Child Language*, *24*(3), 535-565.

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy*, *7*(1), 53-71.

Tomasello, M. (2003). Constructing a language: A usage-based theory of language acquisition. Cambridge, MA, US: Harvard University Press.

Trainor, L. J., & Desjardins, R. N. (2002). Pitch characteristics of infant-directed speech affect infants' ability to discriminate vowels. *Psychonomic Bulletin & Review*, *9*(2), 335-340.
Waxman, S. R., and Gelman, S. A. (2009). Early word-learning entails reference, not merely associations. *Trends Cogn. Sci.* 13, 258–263. doi: 10.1016/j.tics.2009.03.006
Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. *Proc. IEEE* 78, 1550–1560. doi: 10.1109/5.58337
Williams, R. J., and Peng, J. (1990). An efficient gradient-based algorithm for on-line training of recurrent network trajectories. *Neural Comput.* 2, 490–501. doi: 10.1162/neco.1990.2.4.490
Wojcik, E. H., and Saffran, J. R. (2013). The ontogeny of lexical networks: toddlers encode the relationships among referents when learning novel words. *Psychol. Sci.* 24, 1898–1905. doi: 10.1177/0956797613478198
Xie, X., & Seung, H. S. (2003). Equivalence of backpropagation and contrastive Hebbian learning in a layered network. *Neural computation*, *15*(2), 441-454.
Younger, B. (1990). Infant categorization: Memory for category-level and specific item information. *Journal of Experimental Child Psychology*, *50*(1), 131-155.
Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language*, *47*(1), 1-29.
Zipser, D., & Andersen, R. A. (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, *331*(6158), 679.

APPENDIX A: TEST MATERIALS

Probe words and their semantic category

BATHROOM	BIRD	BODY	CLOTHES	DAY	DESSERT	DRINK	ELECTRONIC	FAMILY	FRUIT	MAMMAL
bathtub	chick	arm	apron	friday	cake	beer	alarm	aunt	apple	armadillo
brush	crow	back	bathrobe	monday	candy	cocoa	battery	babe	avocado	bear
comb	cuckoo	beard	belt	saturday	chocolate	coffee	camera	boy	banana	beaver
kleenex	duck	behind	bib	sunday	cookie	coke	clock	brother	berry	buck
lipstick	duckling	belly	blouse	thursday	cream	juice	computer	child	blueberry	buffalo
lotion	eagle	blood	boot	today	cupcake	lemonade	microphone	cousin	cherry	bull
pee	flamingo	bone	buckle	tomorrow	dessert	milk	phone	dad	coconut	bunny
poo	goose	bottom	cap	tonight	fudge	pop	radio	daughter	cranberry	camel
poop	hen	cheek	cape	tuesday	lollipop	soda	register	father	grape	cat
potty	ostrich	chest	coat	wednesday	pie	tea	telephone	girl	grapefruit	cow
shampoo	owl	chin	diaper	week	popsicle	wine	telescope	grandma	lemon	deer
shit	parrot	ear	dress	weekend	pudding	**TIME**	television	grandmother	orange	dingo
shower	peacock	elbow	glove	yesterday	sweet	afternoon	video	grandpa	peach	dog
soap	penguin	eye	hat	**NUMBER**	tapioca	hour	**TOY**	kid	pear	dolphin
sponge	rooster	face	helmet	eight	treat	minute	ball	ma	pineapple	donkey
tissue	**INSTRUMENT**	finger	hood	eighteen	**PLANT**	morning	balloon	mama	plum	elephant
toilet	banjo	foot	jacket	eleven	acorn	night	bat	mom	raisin	fox
toothbrush	bell	forehead	mitten	fifteen	bush	o'clock	block	mother	raspberry	giraffe
toothpaste	drum	hair	outfit	fifty	daisy	second	book	pa	strawberry	goat
towel	flute	hand	pajama	five	flower	**HOUSE**	crayon	papa	tomato	gorilla
tub	guitar	head	purse	forty	grass	backyard	die	pet	watermelon	hamster
KITCHEN	music	heart	scarf	four	lily	basement	doll	sister	**FURNITURE**	hippo
bowl	piano	hip	shirt	fourteen	pine	bathroom	gift	son	bed	horse
cup	tuba	knee	shoe	hundred	seaweed	bedroom	kite	stepmother	bench	hyena
dish	violin	lap	short	nine	tree	ceiling	lego	**VEHICLE**	blanket	jaguar
drain	xylophone	leg	skirt	nineteen	violet	chimney	playdoh	airplane	bookcase	kangaroo
fork	**VEGETABLE**	lip	slipper	one	**TOOL**	closet	puppet	ambulance	cabinet	kitten
freezer	broccoli	memory	sock	seven	broom	counter	puzzle	bicycle	candle	koala
fridge	cabbage	mind	suit	seventeen	drill	curtain	racket	bike	carpet	lamb
glass	carrot	mood	sweater	six	hammer	door	rattle	boat	chair	leopard
knife	celery	mouth	tie	sixteen	iron	driveway	seesaw	bulldozer	couch	lion
microwave	lettuce	mustache	underwear	ten	ladder	fence	sled	bus	crib	monkey
mixer	mushroom	neck	vest	thirteen	lawnmower	floor	sticker	caboose	desk	moose
napkin	olive	nose	**SHAPE**	thirty	mop	garage	teddy	carriage	drawer	mouse
oven	pea	penis	circle	thousand	needle	kitchen	tennis	cart	dresser	opossum
pan	pepper	ponytail	cone	three	pail	nursery	tricycle	dumptruck	dryer	panda
pitcher	pickle	shoulder	cube	twelve	paint	porch	**MEAT**	helicopter	lamp	pig
plate	potato	skin	diamond	twenty	pen	roof	bacon	jeep	pillow	pony
refrigerator	pumpkin	stomach	line	two	pencil	room	beef	jet	seat	porcupine
saucer	salad	throat	loop	zero	rake	sandbox	bologna	motorcycle	shelf	pup
silverware	spinach	thumb	rectangle	**MONTH**	saw	step	fish	stroller	sofa	rabbit
spoon	**INSECT**	toe	square	april	screw	study	flounder	subway	stool	raccoon
stove	ant	tongue	triangle	august	screwdriver	wall	ham	taxi	table	rat
teapot	bee	tooth	**SPACE**	december	shovel	window	hamburger	tractor	wastebasket	reindeer
toaster	butterfly	tummy	earth	january	tape	yard	meat	trailer	**WEATHER**	seal
	caterpillar	weener	jupiter	july	umbrella		roast	train	cloud	sheep
	cricket	wrist	moon	june	vacuum		salmon	truck	fog	skunk
	fly		planet	march	wheelbarrow		steak	van	ice	squirrel
	grasshopper		pluto	may	wrench		tuna	wagon	rain	tiger
	ladybug		star	november					rainbow	walrus
	snail		sun	october					snow	weasel
	spider		world	september					storm	whale
	worm								thunder	wolf
									wind	zebra